YouTube

VIDEO INFORMATION EVALUATION WEB APP

Chien Yuan Chang, Data Scientist

5 BILLION VIDEOS

are watched on Youtube every single day

300 HOURS

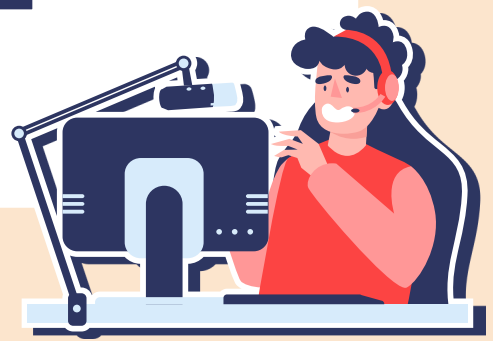
of video are uploaded every minute

$3-$5

are earned per 1000 video views

*Source: https://fortunelords.com/youtube-statistics/
https://influencermarketinghub.com/how-much-do-youtubers-make/

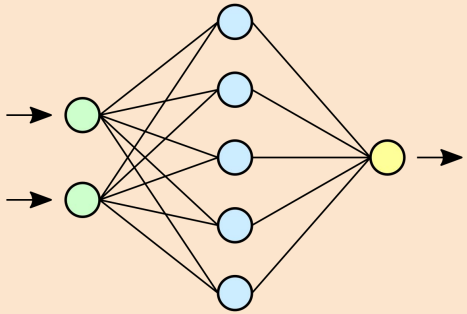
MORE VIEWS
MORE REVENUE

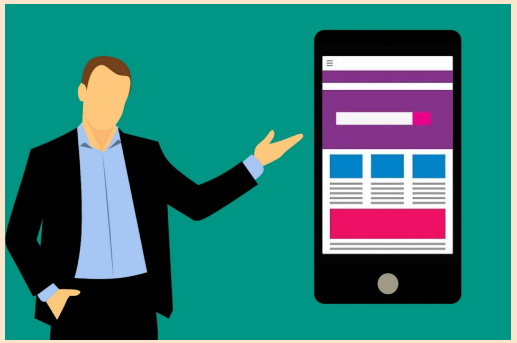


Building a Web App to Estimate the Views can help YouTube Creators optimize Thumbnails and Titles

WORKFLOW

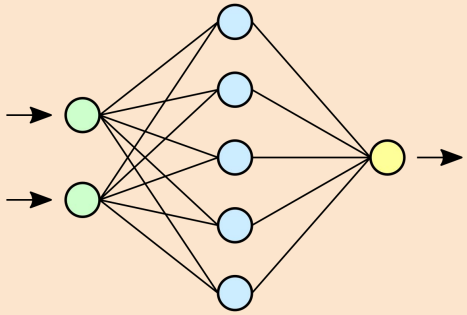
Data Ingestion and Storage → Modeling → Deployment

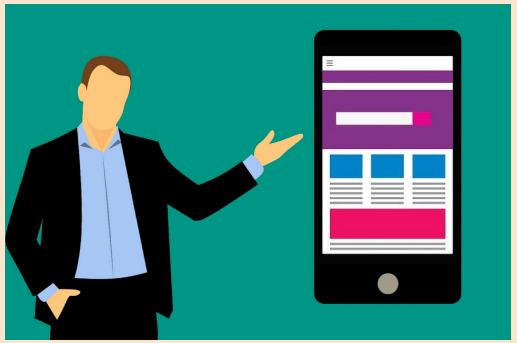




WORKFLOW

Data Ingestion and Storage → Modeling → Deployment





YOUTUBE DATA - FIVE BRANCHES

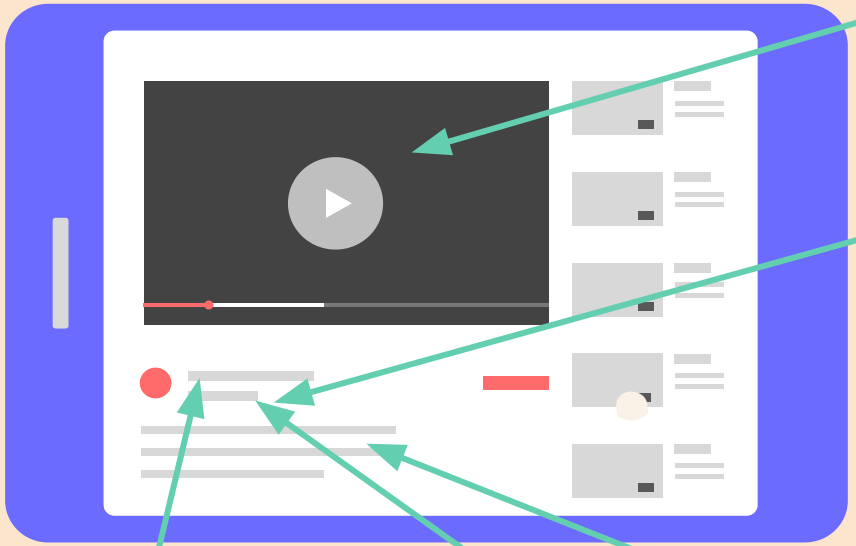


Image Data: Thumbnail

Numerical and Categorical Data:
View Count, Published Time, Duration, Category, Definition, Dimension, Made for kids, Subscriber Count, Channel Country and Channel video Count

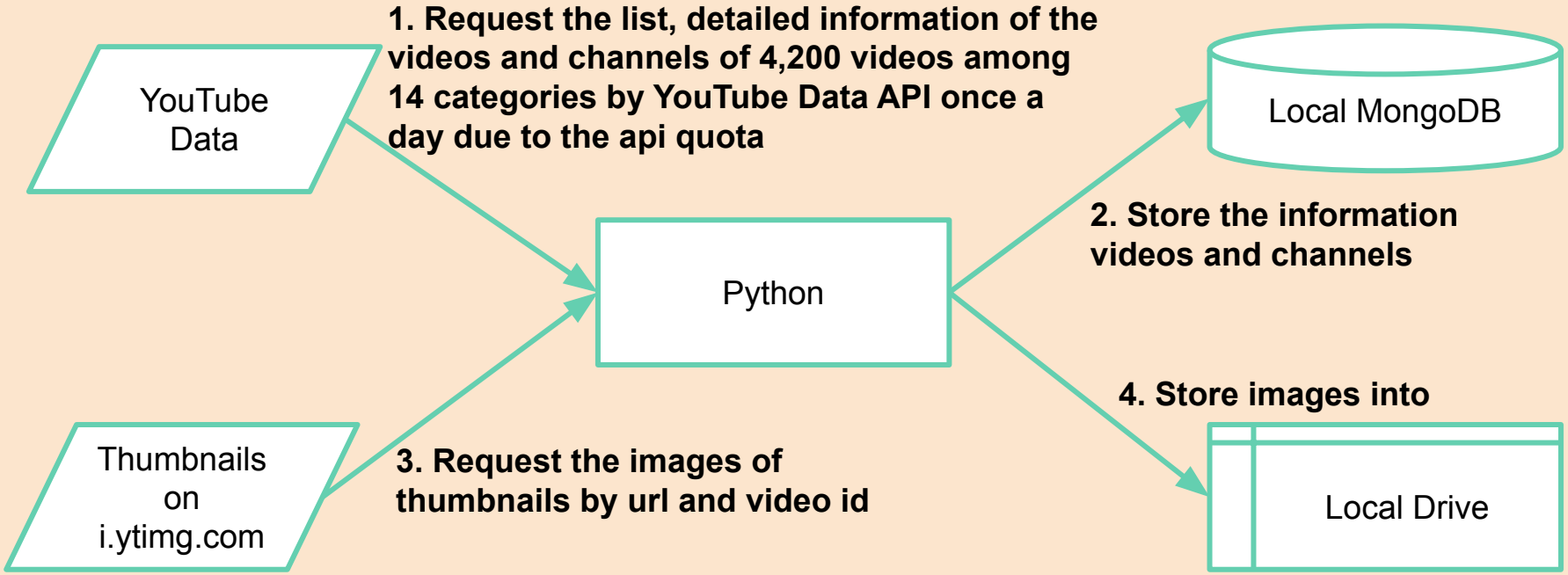
Target:
Log_{10} of Daily Views of the Video

Text Data: Title

Text Data: Tags

Text Data: Description

DATA FLOW DIAGRAM - LOCAL



DATA FLOW DIAGRAM - CLOUD

YouTube Data

Request the list and the detailed information of the videos and channels of 1,400 videos among 14 categories by YouTube Data API automatically every 8 hours due to the api quota

Amazon Lambda (Python)

Send and Store

MongoDB Atlas

Retrieve

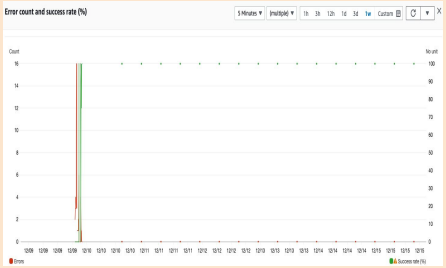
Google Colab (Python)

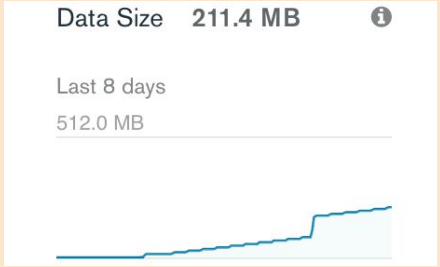
Thumbnails on i.ytimg.com

Request the images of thumbnails by url and video id before training models

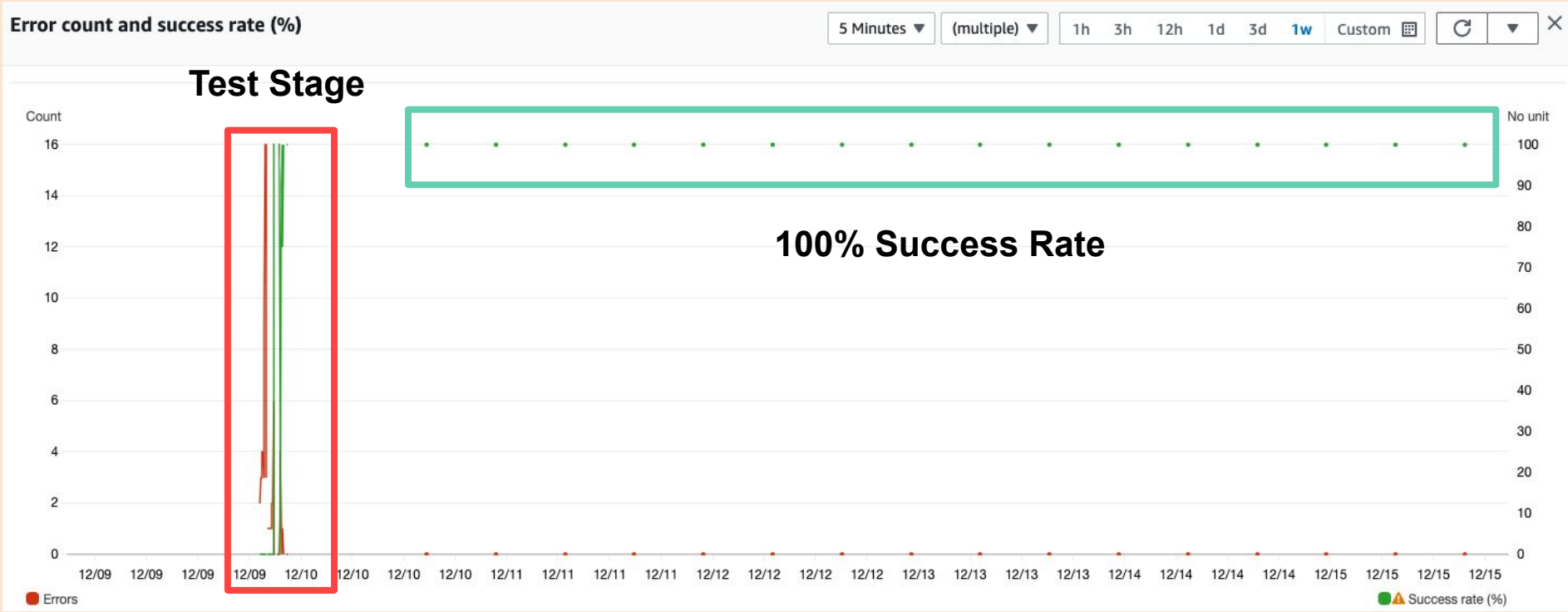
Store images

Google Drive

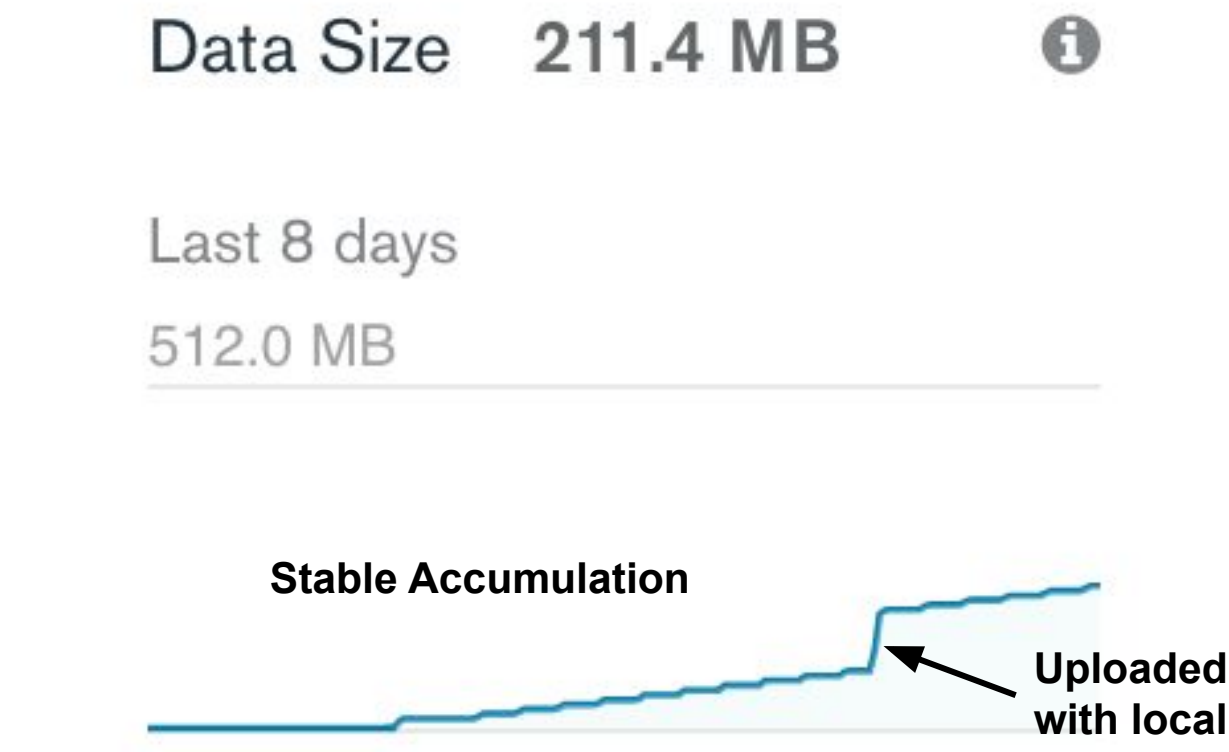




AMAZON LAMBDA - SUCCESS RATE

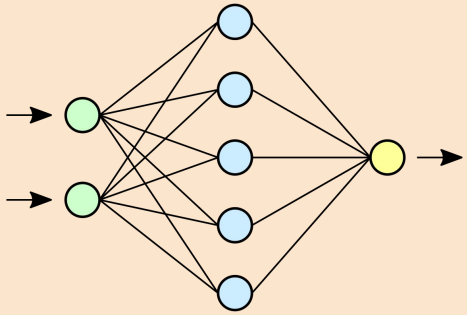


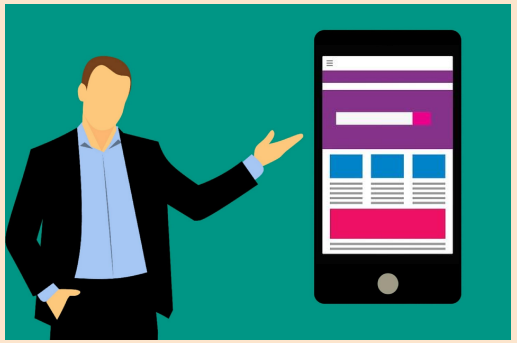
MONGODB ATLAS



WORKFLOW

Data Ingestion and Storage → Modeling → Deployment





DATA FOR MODEL TRAINING

- 18,763 data points

- Total 38,221 data points
- 10,472 unduplicated and clean data points
  - 6,282(60%): Training
  - 2,095(20%): Validation
  - 2,095(20%): Test

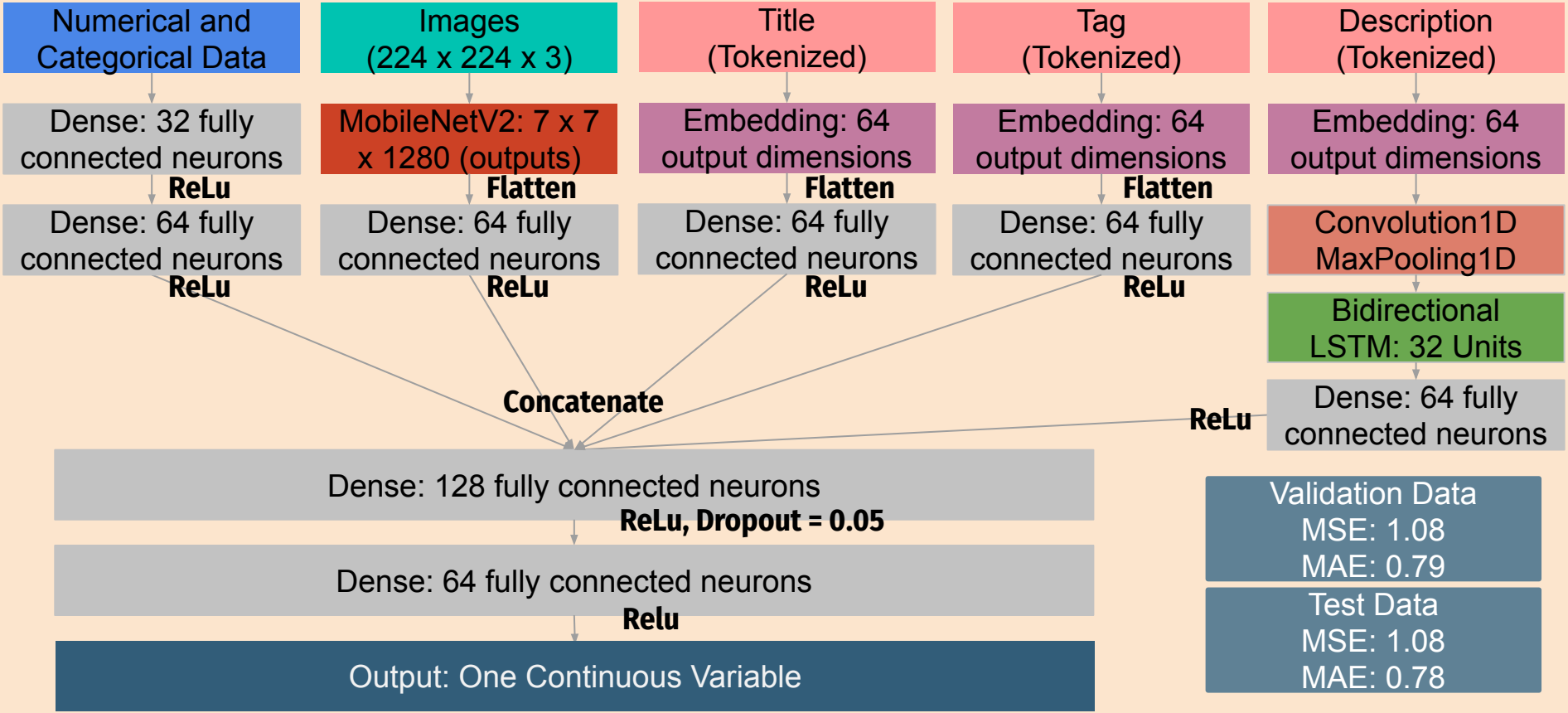
BASELINE MODEL - LINEAR REGRESSION

Feature Branches
Numerical and Categorical Video Information → One-hot Encoding and Standardization
Thumbnails → Resize, Rescale and SVD
Title → NLP, Tokenizer and PCA
Tag → NLP, Tokenizer and PCA
Description → NLP, Tokenizer and PCA

Target:
Log_{10} of Daily Views of the Video

Validation Data
MSE > 1.57
MAE > 0.96

FINAL MODEL - NEURAL NETWORK



Validation Data	
MSE:	1.08
MAE:	0.79

Test Data	
MSE:	1.08
MAE:	0.78

FINAL RESULT FOR THE USERS - COMPARISON BETWEEN USER INPUTS AND BLANK INPUTS

Numerical and Categorical Data	Thumbnail	Title	Tags	Description
User Input	User Input	User Input	User Input	User Input
User Input	Blank Image	Empty	Empty	Empty

Fixed

Prediction

Baseline

Improvement Rate

How many percentage of views can increase when compared with blank image, title, tag and description

FINAL RESULT FOR THE USERS - COMPARISON BETWEEN TWO USER INPUTS

Numerical and Categorical Data	Thumbnail	Title	Tags	Description
Fixed				
User Input	User Input A	User Input A	User Input A	User Input A
User Input	User Input B	User Input B	User Input B	User Input B

Example A

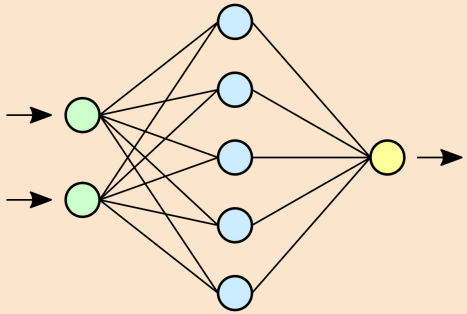
Example B

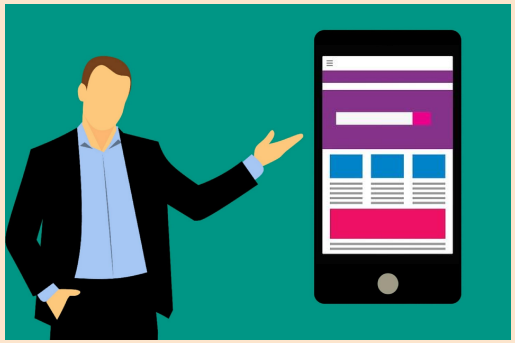
Comparison

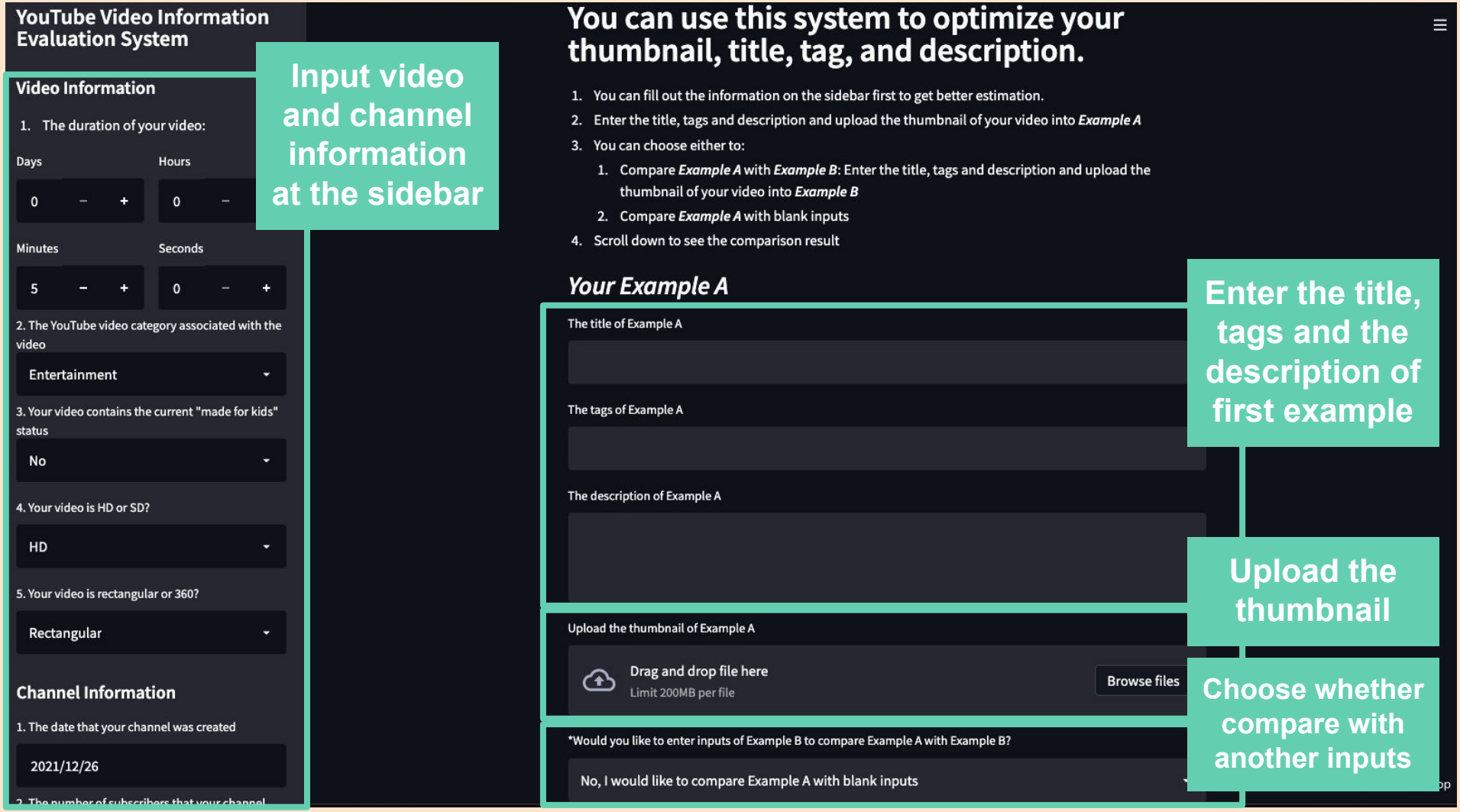
The increasing percentage of views when comparing Example A with Example B.

WORKFLOW

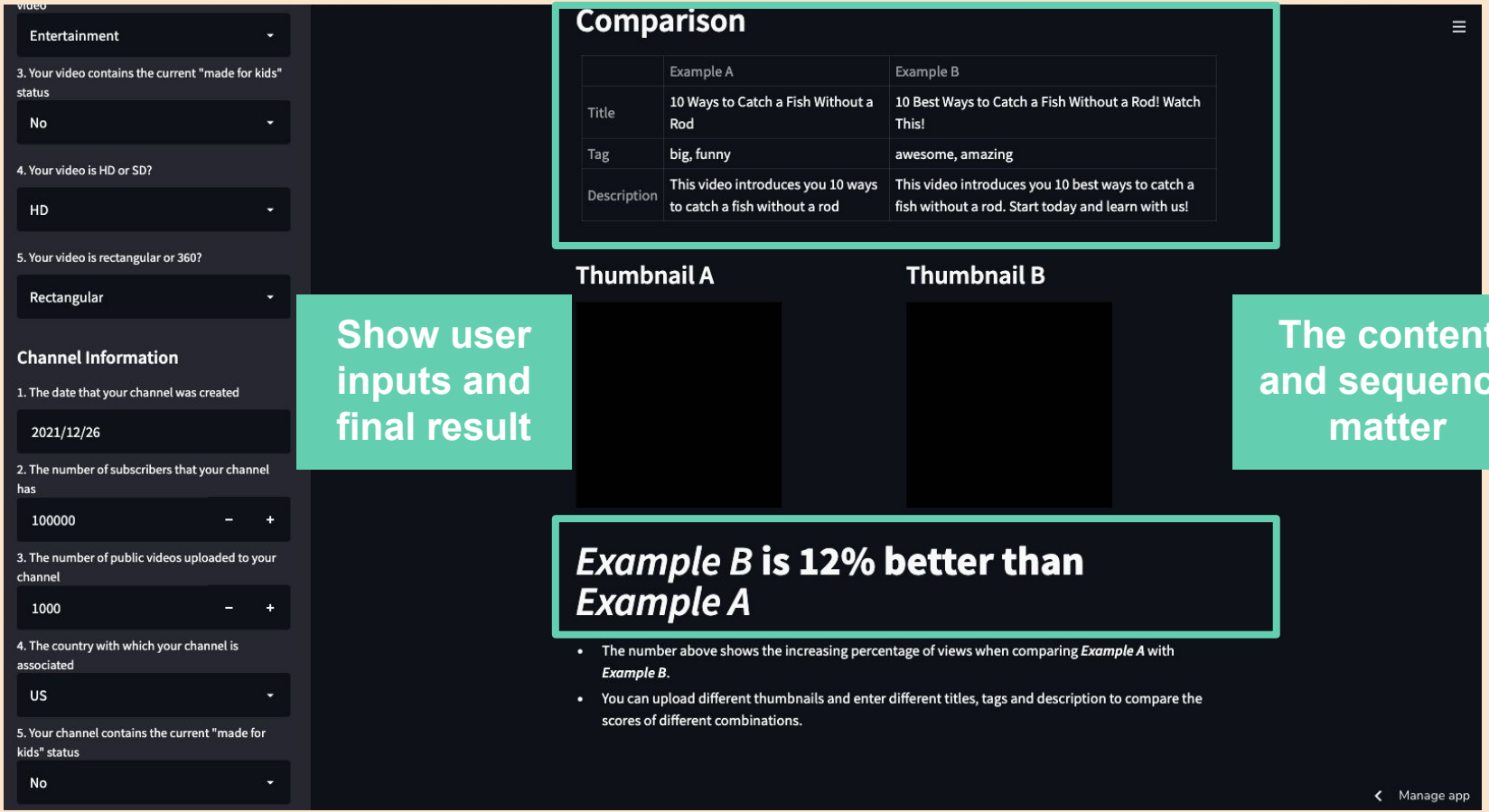
Data Ingestion and Storage → Modeling → Deployment






YouTube Video Information Evaluation System
Input video and channel information at the sidebar
Video Information
1. The duration of your video:
Days
Hours
0
0
Minutes
Seconds
5
0
2. The YouTube video category associated with the video
Entertainment
3. Your video contains the current "made for kids" status
No
4. Your video is HD or SD?
HD
5. Your video is rectangular or 360?
Rectangular
Channel Information
1. The date that your channel was created
2021/12/26
You can use this system to optimize your thumbnail, title, tag, and description.
1. You can fill out the information on the sidebar first to get better estimation.
2. Enter the title, tags and description and upload the thumbnail of your video into Example A
3. You can choose either to:
1. Compare Example A with Example B: Enter the title, tags and description and upload the thumbnail of your video into Example B
2. Compare Example A with blank inputs
4. Scroll down to see the comparison result
Your Example A
The title of Example A
The tags of Example A
The description of Example A
Enter the title, tags and the description of first example
Upload the thumbnail of Example A
Drag and drop file here
Limit 200MB per file
Browse files
Upload the thumbnail
*Would you like to enter inputs of Example B to compare Example A with Example B?
No, I would like to compare Example A with blank inputs
Choose whether compare with another inputs

COMPARE THE TEXT



Show user inputs and final result

The content and sequence matter

COMPARE THE THUMBNAILS

Entertainment

3. Your video contains the current "made for kids" status

No

4. Your video is HD or SD?

HD

5. Your video is rectangular or 360?

Rectangular

Channel Information

1. The date that your channel was created

2021/12/26

2. The number of subscribers that your channel has

100000 − +

3. The number of public videos uploaded to your channel

1000 − +

4. The country with which your channel is associated

US

5. Your channel contains the current "made for kids" status

No

	Example A	Example B
Title		
Tag		
Description		

Thumbnail A

Thumbnail B

Thumbnails make huge impact!

Example B* is 511% better than *Example A

- The number above shows the increasing percentage of views when comparing ***Example A*** with ***Example B***.
- You can upload different thumbnails and enter different titles, tags and description to compare the scores of different combinations.

Made with Streamlit

Manage app

CONCLUSION

- Not precise enough to provide the number of prediction to the users
- A reference for users to compare their inputs, especially thumbnails

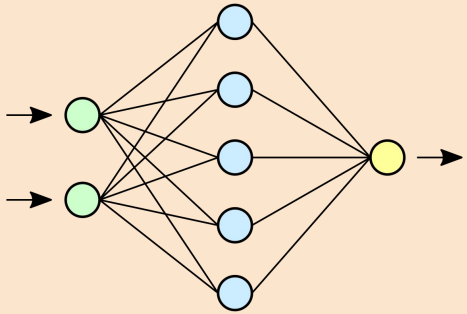
Welcome to try it:

https://share.streamlit.io/koscew/metis_module7_data_engineering_project/main

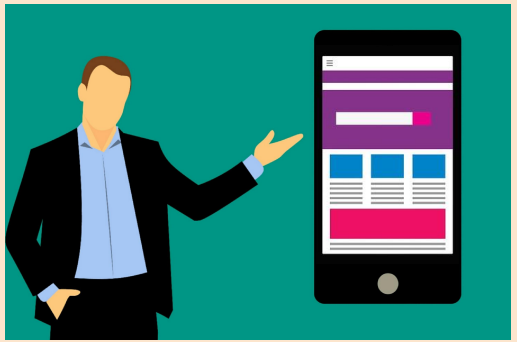
FUTURE WORK

More data

- Increase layers and neurons
- More preprocessing (NLP)
- Time Series

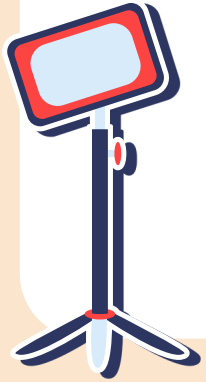


UI Optimization



THANK YOU!

DO YOU HAVE ANY QUESTIONS?



APPENDIX

RESOURCES

- YouTube Data API
  - https://developers.google.com/youtube/v3/docs/
- Amazon Lambda
  - https://aws.amazon.com/lambda/
- MongoDB Atlas
  - https://www.mongodb.com/
- Google Colab
  - https://colab.research.google.com

YOUTUBE DATA API

Search

```
{
  "kind": "youtube#searchResult",
  "etag": etag,
  "id": {
    "kind": string,
    "videoId": string,
    "channelId": string,
    "playlistId": string
  },
  "snippet": {
    "publishedAt": datetime,
    "channelId": string,
    "title": string,
    "description": string,
    "thumbnails": {
      (key): {
        "url": string,
        "width": unsigned integer,
        "height": unsigned integer
      }
    },
    "channelTitle": string,
    "liveBroadcastContent": string
  }
}
```

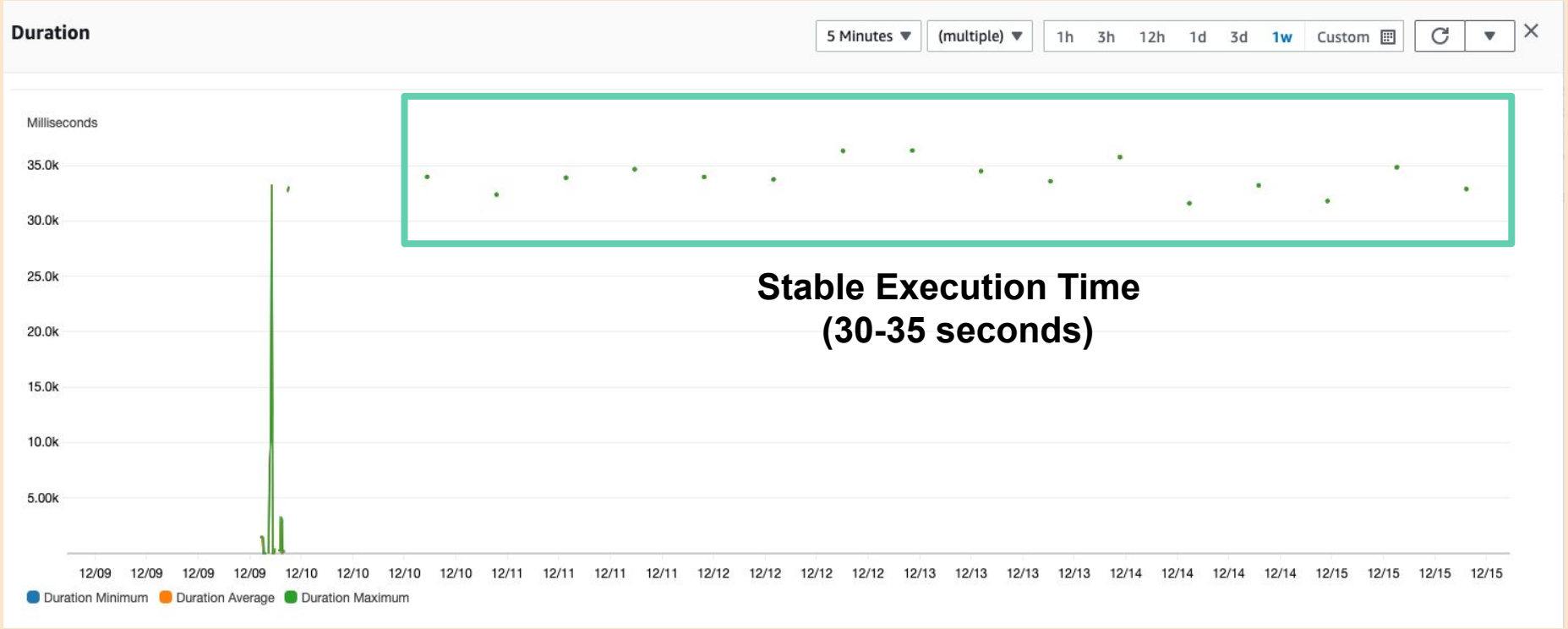
Videos

```
{
  "kind": "youtube#video",
  "etag": etag,
  "id": string,
  "snippet": {
    "publishedAt": datetime,
    "channelId": string,
    "title": string,
    "description": string,
    "thumbnails": {
      (key): {
        "url": string,
        "width": unsigned integer,
        "height": unsigned integer
      }
    },
    "channelTitle": string,
    "tags": [
      string
    ],
    "categoryId": string,
    "liveBroadcastContent": string,
    "defaultLanguage": string,
    "localized": {
      "title": string,
```

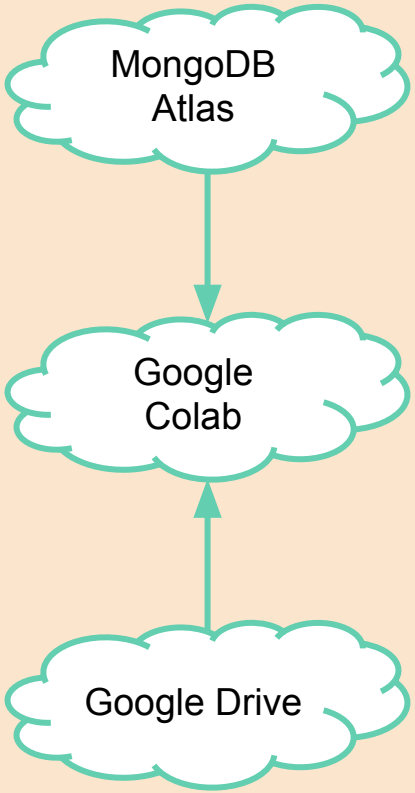
Channels

```
{
  "kind": "youtube#channel",
  "etag": etag,
  "id": string,
  "snippet": {
    "title": string,
    "description": string,
    "customUrl": string,
    "publishedAt": datetime,
    "thumbnails": {
      (key): {
        "url": string,
        "width": unsigned integer,
        "height": unsigned integer
      }
    },
    "defaultLanguage": string,
    "localized": {
      "title": string,
      "description": string
    },
    "country": string
  },
  "contentDetails": {
    "relatedPlaylists": {
```

AMAZON LAMBDA - DURATION



MODELING - NEURAL NETWORK



Feature Branches

Numerical and Categorical Video Information
→ One-hot Encoding and Standardization

Thumbnails
→ Resize, Rescale and Transfer Learning

Title
→ NLP, Tokenizer and Embedding

Tag
→ NLP, Tokenizer and Embedding

Description
→ NLP, Tokenizer, Embedding, CNN and Bidirectional LSTM

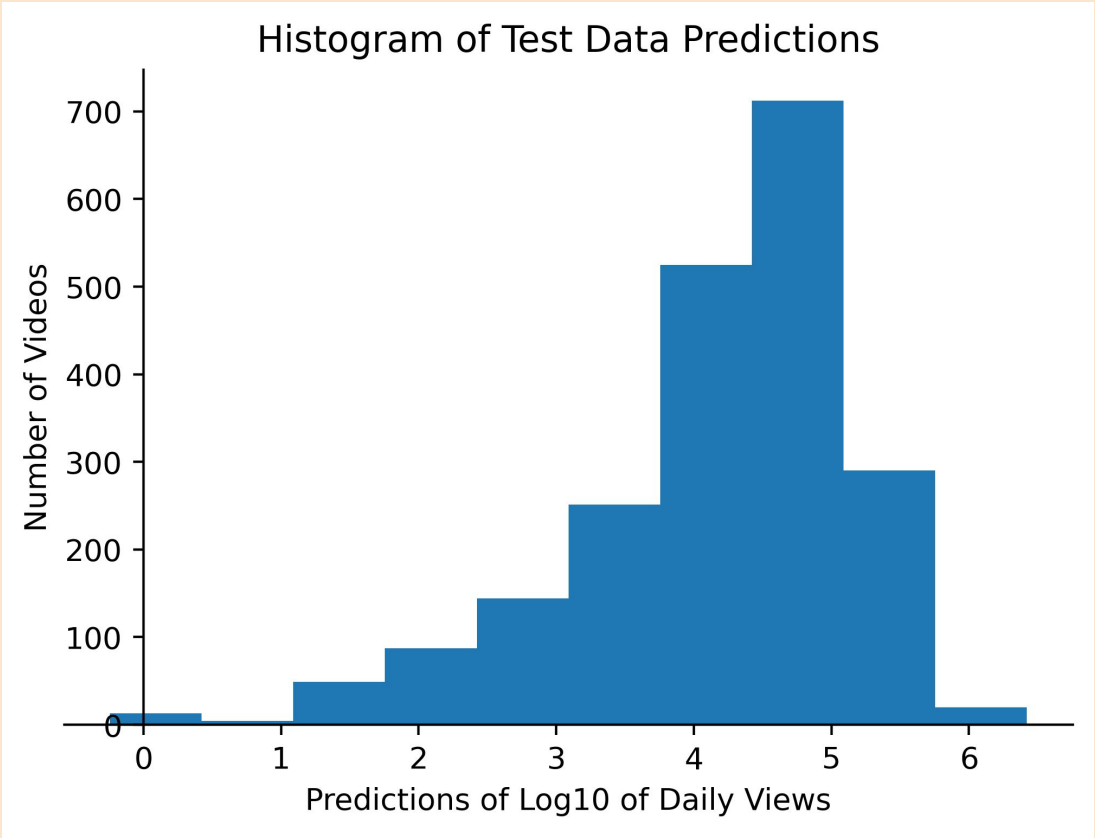
Target:
Log 10 of Daily Views of the Video

Neural Network Model

Validation Data
MSE: 1.08
MAE: 0.79

Test Data
MSE: 1.08
MAE: 0.78

IT CAN BE A REFERENCE



DEPLOY ON STREAMLIT

- Scaler
- Tokenizers
- Model

Streamlit code
on local machine

- Push to Github
- Deploy on streamlit.io