

VidMine: Digging Deep into YouTube Insights

Amulya Ambati
Sneha Sasanapuri
Lakshmi Prasanna Poluru

April 23, 2024

Abstract

With the exponential growth of digital content consumption, understanding user behavior and content trends on platforms like YouTube has become paramount for content creators, marketers, and businesses. "VidMine" presents a comprehensive data mining project aimed at unearthing valuable insights from YouTube's vast repository of data. Leveraging the YouTube Data API and advanced data mining techniques, this project delves deep into YouTube's ecosystem to extract, analyze, and interpret key metrics, trends, and patterns. The project's primary objectives include data extraction, analysis, insight generation, visualization, and recommendations. Through a systematic data collection, "VidMine" uncovers actionable insights pertaining to user engagement, content performance, audience demographics, and emerging trends. Natural Language Processing (NLP) techniques are employed to analyze textual data such as comments and descriptions, providing valuable insights into audience sentiment and engagement.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 3 |
| 2 | Structure of the Report | 3 |
| 3 | Understanding the YouTube Landscape | 3 |
| 4 | Data | 3 |
| 4.1 | Introduction to Data Source | 3 |
| 4.1.1 | Data Retrieval from YouTube Data API | 3 |
| 4.2 | Attributes for analysis | 4 |
| 5 | Methods | 4 |
| 5.1 | Data Cleaning | 4 |
| 5.1.1 | Dealing with Memory issues | 4 |
| 5.1.2 | Parsing Date Column | 5 |
| 6 | Exploratory Data Analysis(EDA) | 5 |
| 6.1 | Data Visualisations | 5 |
| 6.1.1 | Top channels based on views | 5 |
| 6.1.2 | Insights into the Distribution of Content Creation Time Among Channels | 6 |
| 6.1.3 | Analyzing the relationship between Video Duration and View Count | 7 |
| 6.1.4 | Exploring videos durations Across Years: A Box Plot Analysis | 8 |
| 6.1.5 | Distribution of Video Generation on a Monthly Basis | 8 |
| 6.1.6 | Exploring the Relationship Between Likes and Comments: Insights into Viewer Engagement Patterns | 9 |
| 6.1.7 | Word Cloud of videos descriptions | 10 |
| 6.1.8 | Analyzing Viewer Engagement Based on Captions Presence in Videos | 11 |
| 6.1.9 | Exploring Video Duration Distribution Over Time with Captions Analysis | 12 |
| 6.1.10 | Determining association between Duration and Number of Likes | 12 |
| 6.1.11 | Exploring Univariate Distribution of Video Durations | 13 |
| 7 | Implementation of Models | 14 |
| 7.1 | Sentiment Analysis of comments | 14 |
| 7.2 | Random Forest for predicting number of views | 15 |
| 7.3 | LGBM Regression of number of views | 16 |
| 7.4 | XGBM Regression of number of views | 17 |
| 7.5 | Comparison of Model performance | 18 |
| 7.5.1 | Interpretation of Metrics | 18 |

1 Introduction

In the digital age, where online video consumption is at its peak, platforms like YouTube have emerged as ubiquitous sources of entertainment, education, and information. With billions of users and an ever-expanding library of content, YouTube offers a treasure trove of data waiting to be explored. Understanding the dynamics of this vast ecosystem is essential for content creators, marketers, and businesses striving to thrive in the digital landscape.

Through advanced data mining techniques, we embark on a quest to dig deep into the wealth of insights hidden within YouTube’s vast repository of content, engagement metrics, and user interactions. In this report, we present the culmination of our efforts—a comprehensive exploration of YouTube’s data terrain. From data extraction to analysis, visualization, and actionable insights, “VidMine” offers a holistic view of YouTube’s ecosystem and its implications for content creators, marketers, and stakeholders.

2 Structure of the Report

This report is structured to provide a comprehensive overview of the “VidMine” project, encompassing its methodology, findings, insights, and recommendations. We begin by outlining the methodology employed in data extraction, analysis, and visualization. Subsequently, we delve into the insights derived from the analysis, shedding light on user behavior, content trends, and audience engagement metrics. Finally, we conclude with actionable recommendations for leveraging these insights to enhance content strategy, optimize audience engagement, and drive growth on YouTube.

3 Understanding the YouTube Landscape

This paper researches the influence of venue dynamics on cricket match outcomes and prompts the investigation of whether playing at home holds a substantial impact on a team’s likelihood of securing victory, with a consideration of potential exceptions. It also investigates to see if there is an increased likelihood of winning the match for teams that win the toss. Additionally, we aim to discern patterns surrounding teams that demonstrate exceptional prowess or face challenges when competing in away matches.

4 Data

4.1 Introduction to Data Source

Data Retrieval from YouTube Data API

In order to analyze the impact and engagement surrounding the topic of keywords entered in search bar on YouTube, we utilized the YouTube Data API to access various resources such as activities, videos, search, and comments. Through these API endpoints, we were able to gather a comprehensive dataset containing valuable information pertaining to videos related to the specified keyword. For our project, we have taken keywords as “neural networks” and fetched data of top 50 channels.

4.1.1 Data Retrieval from YouTube Data API

Upon querying the YouTube Data API with the search term “neural networks,” we retrieved a multitude of videos matching the search criteria. From these search results, we extracted pertinent information to construct our dataset. The YouTube Data API provides resources to extract data under different module. It is organised effectively to easily identify the different types of resources that can be retrieved using the API.

A “search” resource result contains information about a YouTube video, channel, or playlist that matches the search parameters specified in an API request. An “activity” resource contains information about an action that a particular channel, or user, has taken on YouTube. The actions reported in activity feeds include rating a video, sharing a video, marking a video as a favorite, uploading a video, and so forth. A “channel” resource contains information about a YouTube channel. A video resource represents a YouTube video and is uniquely identified by an id. A “commentThread” resource contains

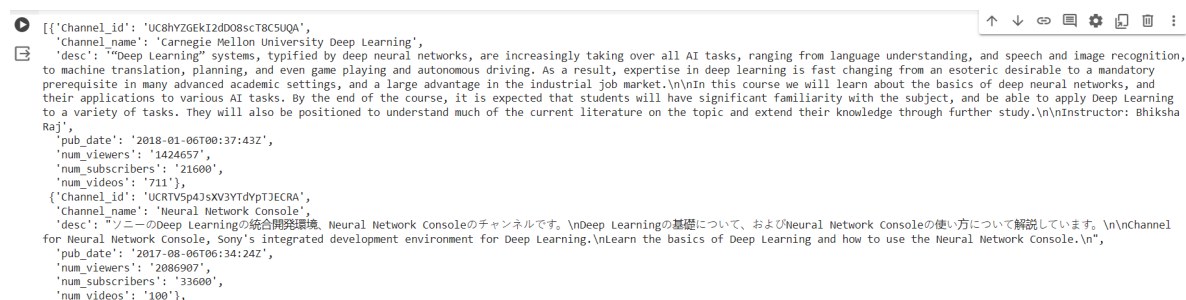
information about a YouTube comment thread, which comprises a top-level comment and replies, if any exist, to that comment. This resource can represent comments about either a video or a channel based on the id supplied.

4.2 Attributes for analysis

The key features extracted include:

- Number of Views: The total number of views garnered by each video, indicating its popularity and reach among viewers.
- Number of Comments: The count of comments posted on each video, providing insights into the level of engagement and interaction generated by the content.
- Date of Published: The timestamp indicating the date and time when each video was published on the platform, enabling temporal analysis and trend identification.
- Number of Likes: The quantity of likes received by each video, reflecting the audience's positive reception and appreciation towards the content.
- Captions Availability: A boolean indicator denoting whether each video has captions or not, facilitating accessibility and catering to diverse audience needs.
- Description: The textual description accompanying each video, offering context and additional information about the content.
- Title: The title of each video, serving as a concise representation of its subject matter and attracting viewer attention.
- Comments: Extracted comments from viewers, providing valuable feedback, opinions, and discussions surrounding the topic of neural networks.
- Time of comment: The extraction of comment timestamps provides valuable insights into the temporal dynamics of user engagement and interaction with YouTube content related to the topic of interest.

The Youtube API returns response in JSON format. Below is the snippet of data before processing.



```
[{"Channel_id": "UC8hY2GEkI2dD08scT8C5UQA",
  "Channel_name": "Carnegie Mellon University Deep Learning",
  "desc": "\"Deep Learning\" systems, typified by deep neural networks, are increasingly taking over all AI tasks, ranging from language understanding, and speech and image recognition, to machine translation, planning, and even game playing and autonomous driving. As a result, expertise in deep learning is fast changing from an esoteric desirable to a mandatory prerequisite in many advanced academic settings, and a large advantage in the industrial job market.\n\nIn this course we will learn about the basics of deep neural networks, and their applications to various AI tasks. By the end of the course, it is expected that students will have significant familiarity with the subject, and be able to apply Deep Learning to a variety of tasks. They will also be positioned to understand much of the current literature on the topic and extend their knowledge through further study.\n\nInstructor: Bhiksha Raj",
  "pub_date": "2018-01-06T00:37:43Z",
  "num_views": 142457,
  "num_subscribers": 216000,
  "num_videos": 711},
  ...
  {"Channel_id": "UCRTVSp4JsXV3YtdypTJECRA",
  "Channel_name": "Neural Network Console",
  "desc": "\"ソニーのDeep Learningの統合開発環境、Neural Network Consoleのチャンネルです。\\nDeep Learningの基礎について、およびNeural Network Consoleの使い方について解説しています。\\n\\nChannel for Neural Network Console, Sony's integrated development environment for Deep Learning.\\nLearn the basics of Deep Learning and how to use the Neural Network Console.\\n",
  "pub_date": "2017-08-06T06:34:24Z",
  "num_views": 2086907,
  "num_subscribers": 33600,
  "num_videos": 100}]
```

Figure 1: Snapshot of data before processing

5 Methods

5.1 Data Cleaning

5.1.1 Dealing with Memory issues

We have encountered memory issues while reading data of comments into working space. This has made us to read comments data into small chunks of 1000 size at a time and then looping through them

in order to convert into dataframe chunk by chunk. As we have pulled data from API, there is very minimal chances of missing values. Null values are present only in 'desc' column which is categorical in nature and doesn't need any imputation method.

5.1.2 Parsing Date Column

The date columns in the dataset were subjected to parsing and standardization to ensure consistency and facilitate meaningful temporal analysis. This process involved converting the date information from its original format into a standardized date format that is easily interpretable and compatible with analytical tools.

After all steps of data cleaning and processing, the dataframe looks like as follows:

| | Unnamed: 0 | Channel_Id | Video_name | Video_Id | desc | pub_date | num_views | num_likes | num_fav | captions | duration | num_comments | duration_minutes | days_since_pub | pub_month |
|---|------------|--------------------------|--|-------------|---|---------------------------|-----------|-----------|---------|----------|----------|--------------|------------------|----------------|-----------|
| 0 | 0 | UCRTV5p4JxXV3YTdYpTJECRA | Basics of Designing Neural Network - Introd... | Ip4skESU8Cs | This video explains the basics of designing ne... | 2022-09-12 01:10:38+00:00 | 1505 | 9.0 | 0 | True | PT24M33S | 0.0 | 24.55 | 558 days | September |
| 1 | 1 | UCRTV5p4JxXV3YTdYpTJECRA | NVC Tutorial: How to create dataset for image... | gAlzxNCH-KA | In this video, we will explain how to create a... | 2022-09-12 01:10:39+00:00 | 5815 | 43.0 | 0 | True | PT10M33S | 0.0 | 10.55 | 558 days | September |

Figure 2: Snapshot of data after processing

6 Exploratory Data Analysis(EDA)

Our goal is to unearth meaningful stories embedded in the data and set the stage for more targeted analyses. We aim not only to uncover the inherent characteristics of the dataset but also to lay the groundwork for informed decision-making in subsequent phases of our data exploration.

Using the data insights acquired through the analysis of different features, the creation of various info graphics, and the incorporation of our existing knowledge of how Youtube gives search results and what criteria makes it possible.

6.1 Data Visualisations

6.1.1 Top channels based on views

In order to identify the most viewed channels within the realm of content related to our topic of interest [keyword (neural networks)], we conducted an analysis of the total number of views garnered by individual channels. This analysis provides valuable insights into the popularity and reach of content creators within our target domain. **3Blue1Brown** channel emerged as the front runner with a remarkable total view count, reflecting its broad appeal and engaging content offerings. Following closely, **Sebastain Lague** channel and **StatQuest with Josh Starmer** channel demonstrated strong viewer engagement and resonance, securing prominent positions in the ranking. **Deep Learning AI** and **Neural Networks & Deep Learning** channels rounded out the top five with consistent performance and audience affinity, highlighting their relevance and influence within the niche content landscape.

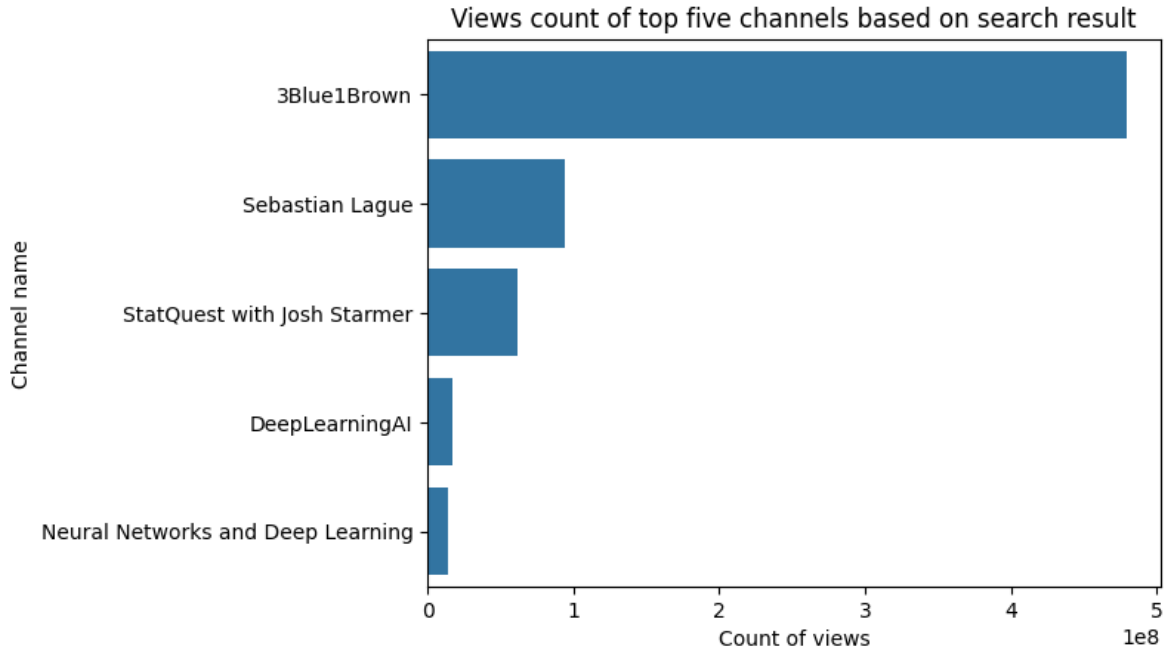


Figure 3: Top Five Channels Based on Number of Views: Insights

6.1.2 Insights into the Distribution of Content Creation Time Among Channels

In this pie chart, we have visualized the top 10 channels by distribution of their content creation time. This visualization provides a clear overview of the proportion of content creation efforts contributed by these channels. This analysis enhances our understanding of the content creation dynamics within our target domain and informs strategic decision-making regarding content partnerships, audience targeting, and engagement initiatives. This pie chart reveals **Alexander Amini** channel as the leading contributor, with 19% of the total content creation minutes, followed closely by **Killian Weinberger** channel with 13%. Remaining channels collectively contribute varying shares, ranging from 5% to 12%.

Percentage distribution of top 10 channels based on minutes of content creation

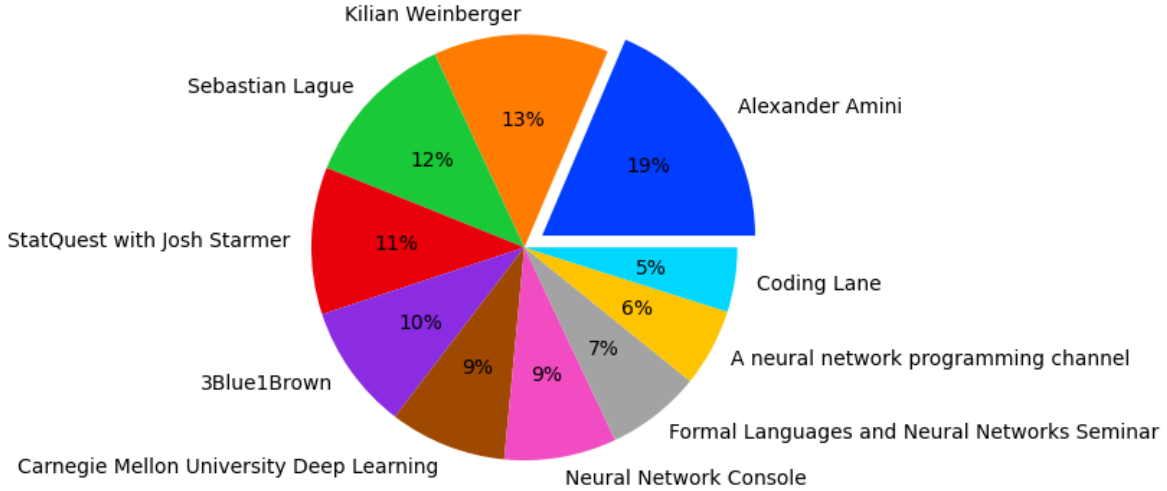


Figure 4: Breakdown of Top 10 Channels by Share of Content Creation Time

6.1.3 Analyzing the relationship between Video Duration and View Count

We utilized Seaborn's regplot function to generate a scatter plot examining the relationship between the duration of videos and the number of views they accrued. The data was filtered to include only videos with a view count of 100,000 or fewer views, ensuring a focused analysis on relatively less-viewed content. From this, we can infer that there is no preference for shorter videos among users which is proved by an almost straight line in the plot. It indicates very weak or no relationship between duration of video length and number of views.

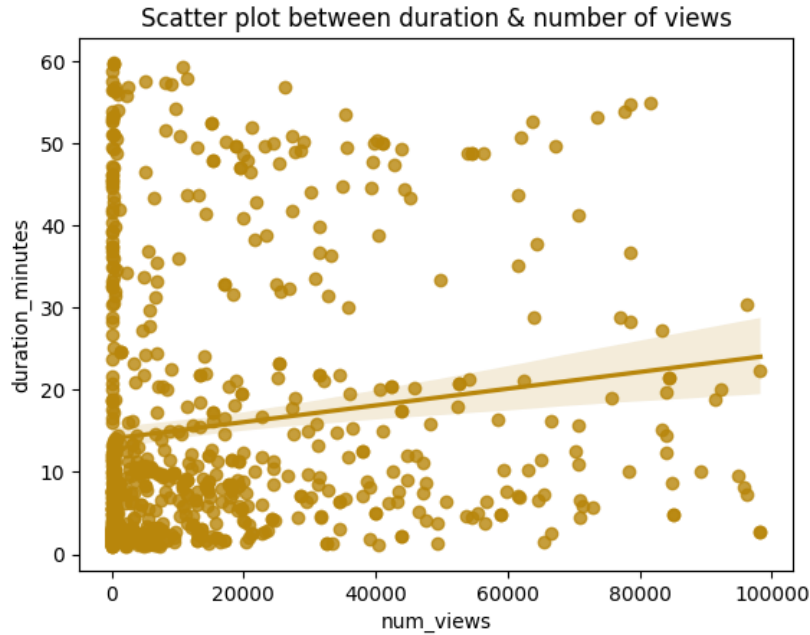


Figure 5: Linear regression between video length & views

6.1.4 Exploring videos durations Across Years: A Box Plot Analysis

In our analysis of the year-wise distribution of video duration, we utilized a box plot visualization to explore how the duration of videos has evolved over time. From the visualization, we observe variations in video duration across different years, with some years exhibiting a wider range of durations than others. It is evident that there is more range in the year **2018** spanning to larger length videos. This analysis provides valuable insights into trends and patterns in video duration over time, informing content creators and stakeholders about audience preferences and content consumption habits across different years.

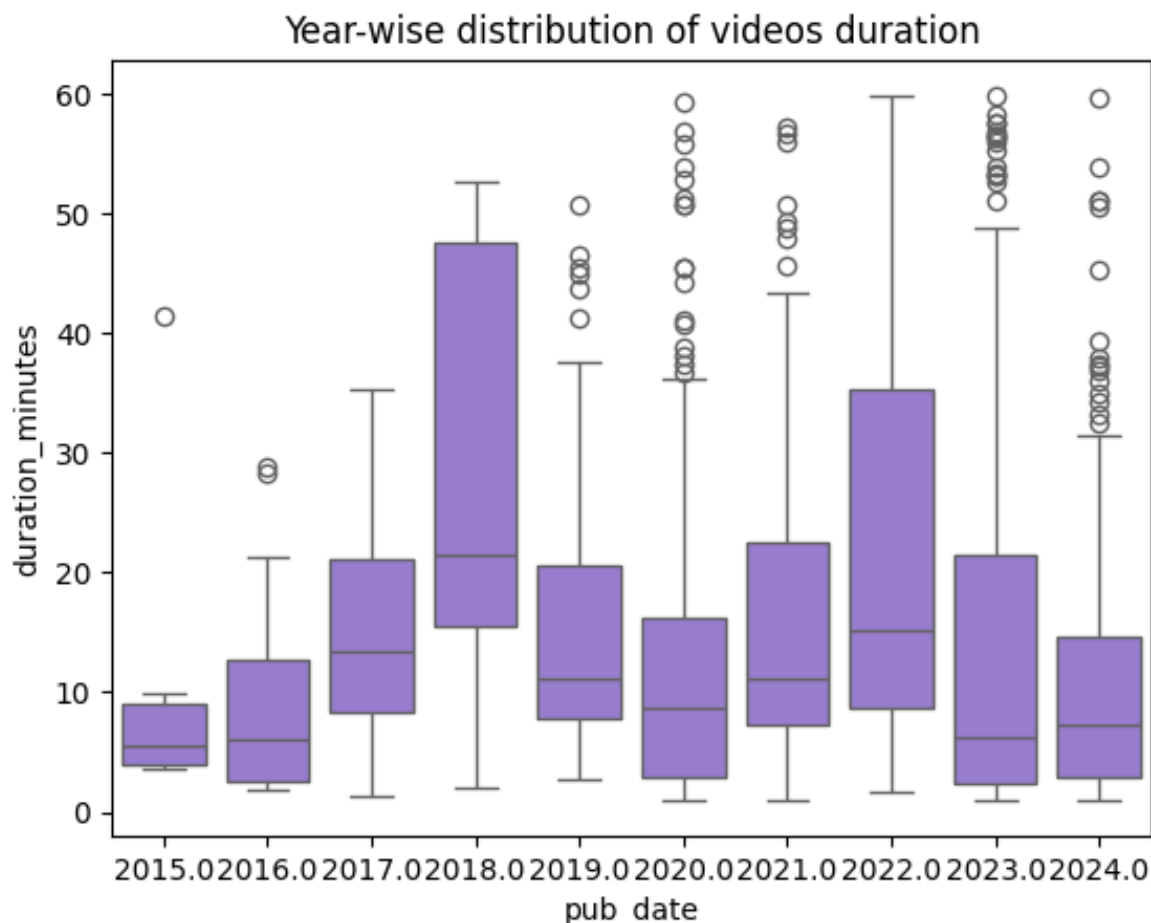


Figure 6: Year-wise Distribution of Video Duration

6.1.5 Distribution of Video Generation on a Monthly Basis

The histogram visualization depicts the month-wise distribution of the number of videos generated within our dataset. We can clearly see that there are more number of videos generated in early months of year compared to mid year months and again picking up pace in later months of the year.

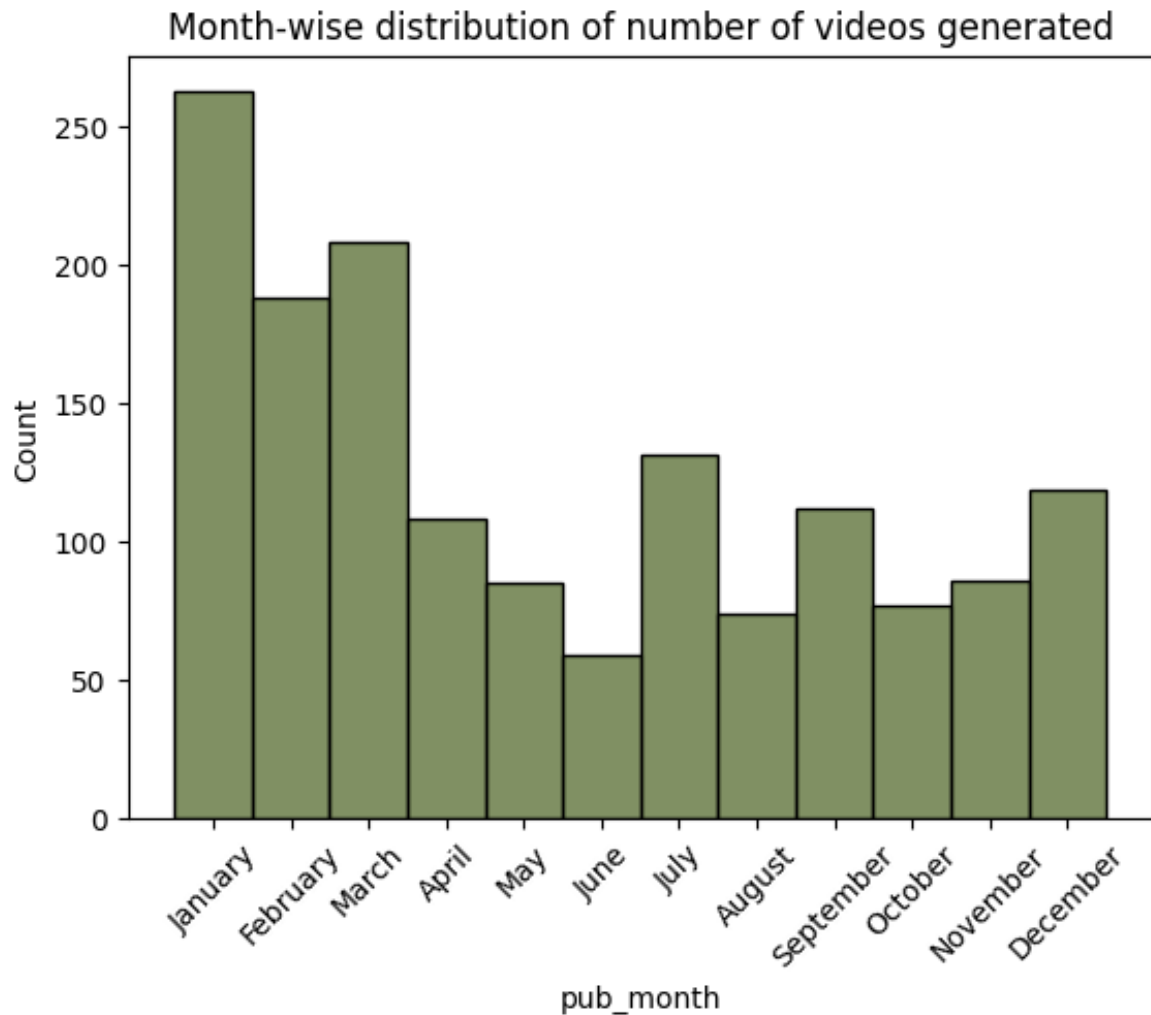


Figure 7: Month-wise distribution of number of videos generated

6.1.6 Exploring the Relationship Between Likes and Comments: Insights into Viewer Engagement Patterns

In our analysis, we utilized a scatter plot visualization as a tool to investigate the correlation between two key engagement metrics: the number of likes and the number of comments garnered by each video within our dataset. While it might seem intuitive that individuals would comment on a video they have watched or liked, the visualization allowed us to delve deeper into the underlying patterns of human behavior. Specifically, we sought to understand the propensity for individuals to engage with content by leaving comments, particularly in instances where they have expressed a positive or negative sentiment through liking or disliking the video. This observation underscores a fundamental aspect of human nature – the tendency to express opinions or reactions when experiencing an emotional response. Consequently, our analysis suggests a symbiotic relationship between the number of comments and likes, where heightened engagement in one metric often corresponds to increased activity in the other.

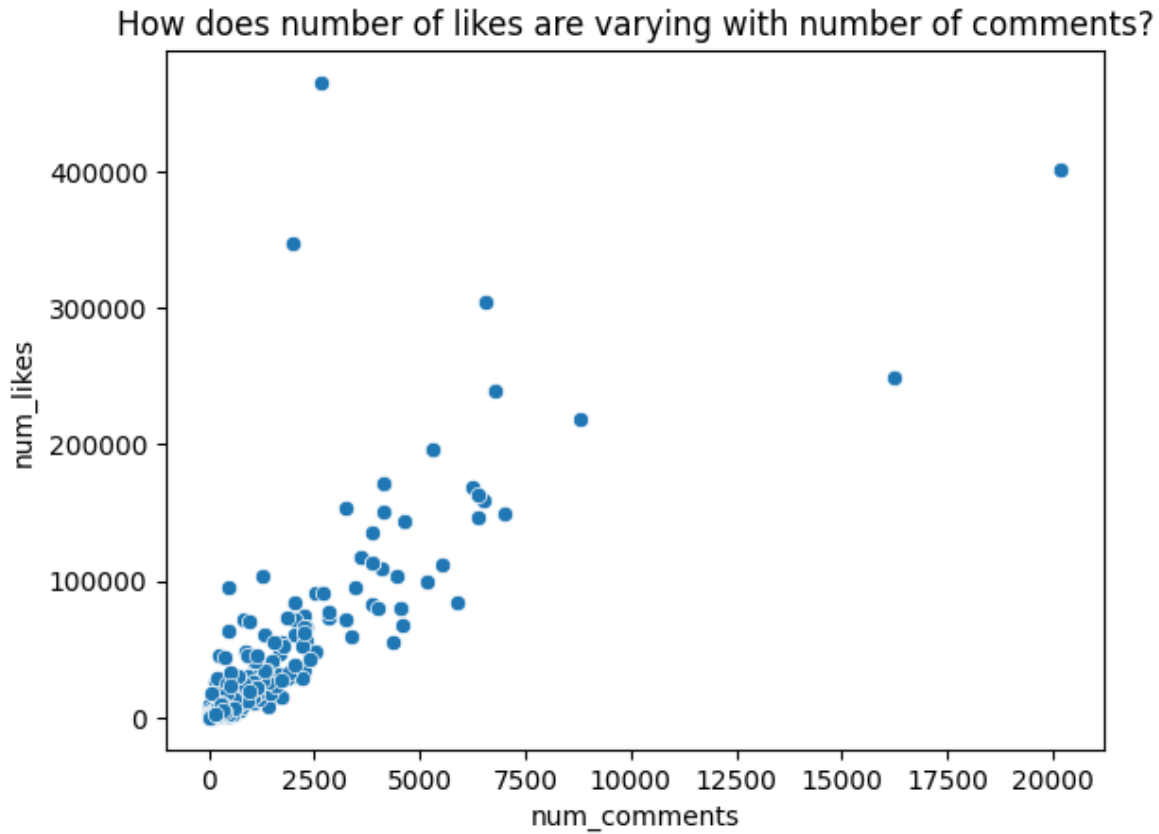


Figure 8: Scatterplot between comments and likes count

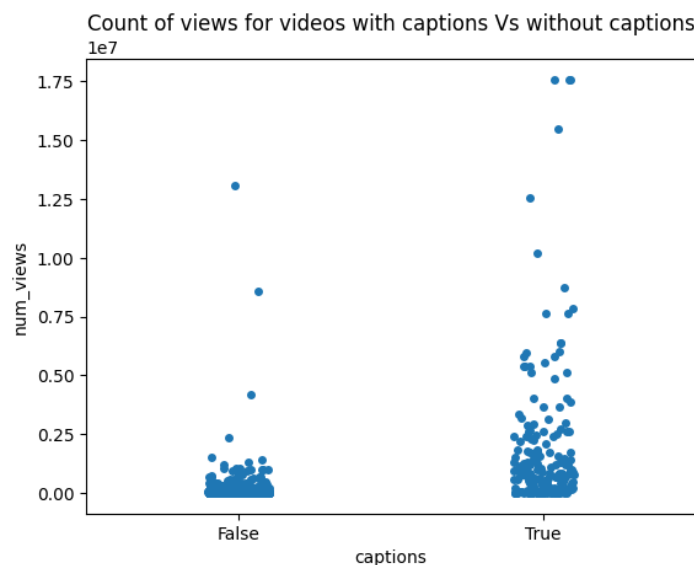
6.1.7 Word Cloud of videos descriptions

Upon analyzing the word cloud generated from the descriptions of videos related to neural networks, several key terms emerged prominently, providing valuable insights into the prevalent themes and topics within this domain. Notably, words such as "machine learning," "deep learning," "beginner," "topic," "AI," and "specialization" recurred frequently, indicating their significance and prevalence in the descriptions of these videos. The prominence of these terms suggests that the videos likely cover introductory concepts and discussions surrounding machine learning and deep learning, catering to individuals with varying levels of expertise, including beginners seeking to explore the topic further. Additionally, the mention of "AI" underscores the intersectionality of neural networks with artificial intelligence, reflecting the broader context in which these videos are situated. Overall, the word cloud provides valuable insights into the dominant themes and focal points of the videos, guiding viewers and content creators alike in navigating the expansive landscape of neural network-related content on the platform.



6.1.8 Analyzing Viewer Engagement Based on Captions Presence in Videos

The strip plot visualization showcases the distribution of views for videos categorized based on the presence or absence of captions. The x-axis represents the captions status, with "Yes" indicating videos with captions and "No" indicating videos without captions, while the y-axis depicts the corresponding count of views. The plot provides a comparative analysis, allowing us to discern any potential differences in viewership between videos with and without captions. The title, "Count of views for videos with captions Vs without captions," succinctly summarizes the main focus of the visualization, facilitating clear interpretation. This visualization serves as a valuable tool for exploring the impact of captions on viewer engagement and accessibility within our dataset of videos, offering insights that can inform content creators and stakeholders in optimizing their video content strategies. We can see that videos having captions have more number of views.



6.1.9 Exploring Video Duration Distribution Over Time with Captions Analysis

This catplot visualization employs violin plots to explore the distribution of video durations across different years, with a distinction made between videos with and without captions. These plots display the distribution of video durations within each year, with the width indicating the density of videos at different duration values. The hue parameter separates the violin plots based on the presence or absence of captions, allowing for a comparative analysis of video durations between the two categories. The mean of duration of videos with captions is relatively higher than those videos without captions. By utilizing violin plots, this visualization effectively illustrates the variability and distribution of video durations over time, while also highlighting any discernible patterns or differences in duration between videos with and without captions.

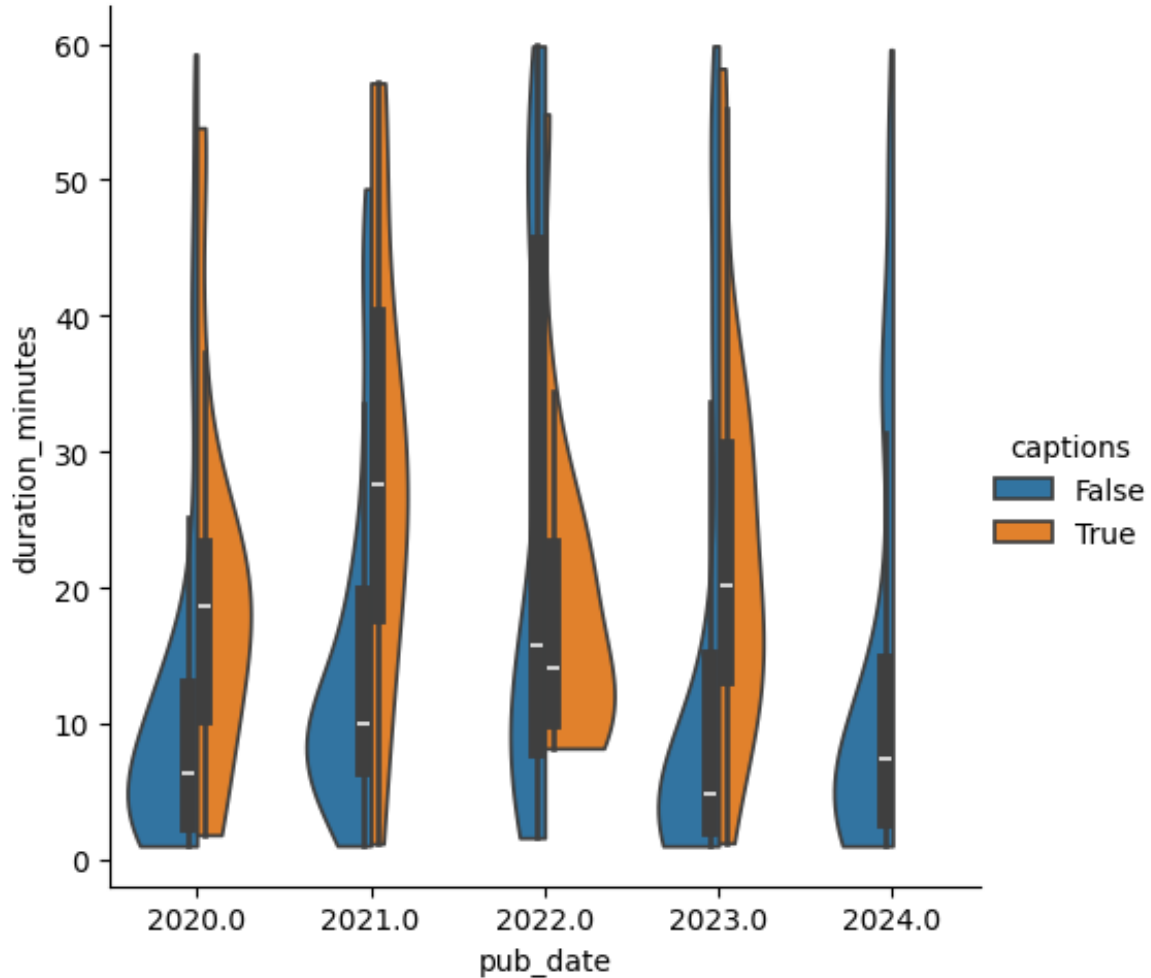


Figure 11: Plot of number of views against caption

6.1.10 Determining association between Duration and Number of Likes

This jointplot reveals that there is no clear conclusion to draw but this plot helps in understanding independence between these two features. Also, it is not very clear to interpret the relationship between these features. This visualization serves as a valuable tool for understanding how video duration influences viewer engagement, as measured by the number of likes, while also considering the presence of captions as a potential factor influencing viewer interaction.

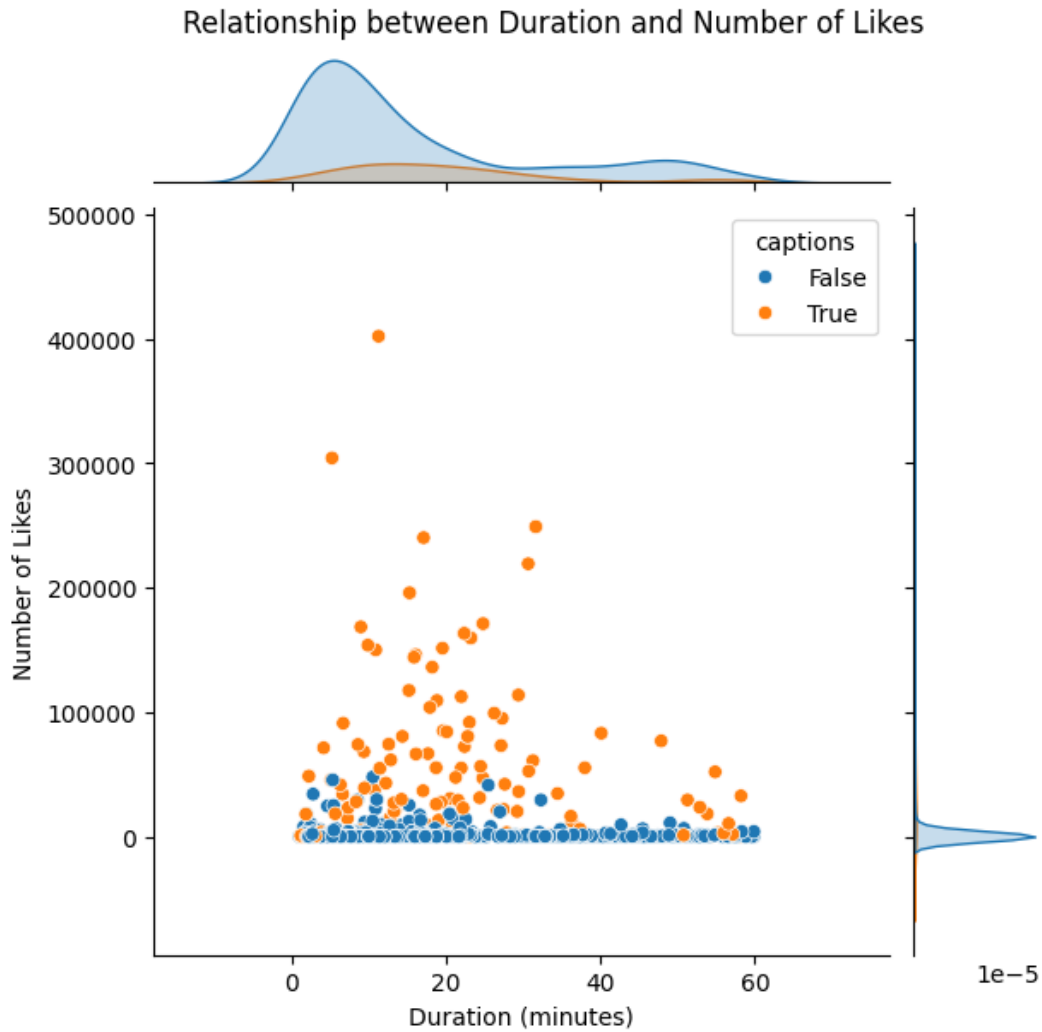


Figure 12: Jointplot of number of likes and duration of videos

6.1.11 Exploring Univariate Distribution of Video Durations

The swarm plot visualization provides a univariate distribution of the duration of videos within our dataset. Each data point on the plot represents an individual video, with the x-axis depicting the duration of the videos in minutes. From the visualization, we observe a wide range of durations, spanning from shorter to longer videos. Additionally, the swarm plot highlights the density of videos at lower duration values, with clusters of data points indicating areas of higher prevalence. Overall, this visualization offers valuable insights into the distribution and variability of video durations within our dataset, informing our understanding of the content landscape and helping to identify potential trends or patterns in video duration.

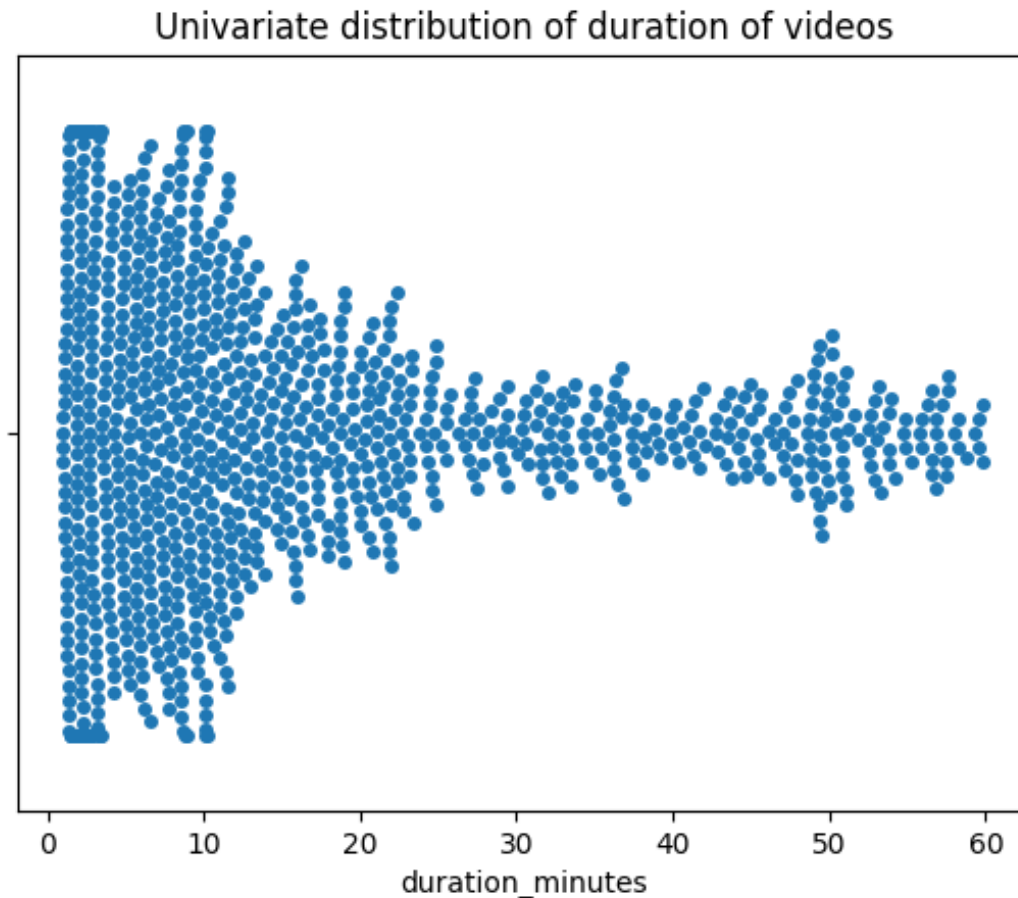


Figure 13: Swarmplot of number of likes and duration of videos

7 Implementation of Models

In this section, we present the implementation of machine learning (ML) models on our YouTube dataset, aimed at uncovering insights into video performance and audience engagement. YouTube, as one of the largest platforms for video content consumption, presents a rich source of data encompassing a myriad of metrics such as views, likes, comments, and subscriber counts. Leveraging this dataset, our analysis seeks to harness the power of ML techniques to predict key performance indicators and understand the factors influencing video success. This sets the stage for our exploration of ML-driven analysis on YouTube data, offering a glimpse into the potential of data-driven approaches in unlocking the platform's dynamics and driving informed decision-making.

7.1 Sentiment Analysis of comments

In conducting sentiment analysis on the comments within our YouTube dataset, we employed natural language processing (NLP) techniques to extract insights into the sentiments expressed by viewers. Initially, we preprocessed the text data by removing comments which are not in English and also the duplicated comments. Subsequently, we utilized VADER (Valence Aware Dictionary and sEntiment Reasoner) library to compute polarity scores for each comment. These polarity scores quantified the sentiment of each comment on a numerical scale, typically ranging from -1 (indicating extremely negative sentiment) to +1 (indicating extremely positive sentiment). We categorized the polarity values as 'positive' if greater than 0.05, 'negative' if less than -0.05, and 'neutral' otherwise. By aggregating these polarity scores across all comments for each video, we gained a comprehensive understanding of the overall sentiment towards the content. This approach to sentiment analysis provided valuable insights into the audience's emotional response to the videos on the YouTube platform, facilitating

informed decision-making for content creators and marketers. Now, let's have a look at snapshot of the dataset before and after transformation illustrating the preprocessing steps undertaken to prepare the data for analysis.

| Unnamed: 0 | channel_id | video_id | comment_text | comment_time |
|------------|------------|--------------------------|---|----------------------|
| 0 | 0 | UCRTV5p4JsXV3YTdYpTJECRA | tFmQj7W4qlk Can you provide a simple example calculation o... | 2023-06-01T11:53:06Z |
| 1 | 1 | UCRTV5p4JsXV3YTdYpTJECRA | tFmQj7W4qlk who chose this music :/ | 2022-12-27T16:17:47Z |
| 2 | 2 | UCRTV5p4JsXV3YTdYpTJECRA | tFmQj7W4qlk Nice work. Just that the background music is s... | 2022-12-02T06:19:11Z |
| 3 | 3 | UCRTV5p4JsXV3YTdYpTJECRA | tFmQj7W4qlk For anyone interested for the music in the bac... | 2022-11-09T08:45:29Z |
| 4 | 4 | UCRTV5p4JsXV3YTdYpTJECRA | tFmQj7W4qlk Can you provide a simple example calculation o... | 2023-06-01T11:53:06Z |

Figure 14: Snapshot of data before analysis

Before transformation, the raw dataset consists of a collection of YouTube video data, containing features such as video titles, descriptions, comments, likes, views, number of favorites, no. of days since published and other relevant metrics. Each row represents a unique video entry, with associated attributes reflecting various aspects of video content and viewer engagement.

| | Channel_id | num_viewers | num_likes | num_fav | captions | duration_minutes | days_since_published | negative | neutral | positive |
|---|--------------------------|-------------|-----------|---------|----------|------------------|----------------------|----------|---------|----------|
| 0 | UCRTV5p4JsXV3YTdYpTJECRA | 4812 | 74.0 | 0 | True | 10.116667 | 589 | 1.0 | 1.0 | 2.0 |
| 1 | UCRTV5p4JsXV3YTdYpTJECRA | 7764 | 41.0 | 0 | False | 20.350000 | 589 | 3.0 | 1.0 | 1.0 |
| 2 | UCRTV5p4JsXV3YTdYpTJECRA | 8290 | 109.0 | 0 | False | 20.083333 | 1068 | NaN | NaN | 1.0 |
| 3 | UCRTV5p4JsXV3YTdYpTJECRA | 8966 | 109.0 | 0 | False | 22.600000 | 1076 | NaN | NaN | 1.0 |
| 4 | UCRTV5p4JsXV3YTdYpTJECRA | 6403 | 117.0 | 0 | False | 3.233333 | 1130 | NaN | 1.0 | NaN |

Figure 15: Snapshot of data after polarity scores

Following transformation, the dataset undergoes several steps to enhance its suitability for analysis. This includes text preprocessing techniques such as removing special characters, stopwords, and tokenization.

7.2 Random Forest for predicting number of views

In our analysis of the YouTube dataset using random forest regression, we aimed to predict the number of views for various videos. Our evaluation metrics provide insights into the model's performance. The R-squared (R^2) score of 0.9 indicates that our model explains approximately 90% of the variability in the number of views, suggesting a strong ability to capture trends and patterns within the data. Despite this, there are opportunities for improvement to reduce prediction errors and enhance the model's accuracy. Future iterations could focus on refining the model by incorporating additional features or exploring alternative regression techniques. Additionally, delving deeper into the dataset and considering factors such as viewer engagement metrics or video content characteristics may offer valuable insights for optimizing strategies to maximize viewership on the YouTube platform. Below is the glimpse of dataset before predicting. This is the same input format which we have used in all our regression models. We have transformed all all categorical columns to one-hot encoding. Features were scaled down to same scale in order to avoid bias of model towards any large-scaled features. Below are the snippets of dataset before and after regression. After regression, the dataset is enriched with predicted values for the number of views, providing insights into the expected viewership for each video based on the learned patterns and relationships within the data. The results of this ML model are tabulated as below. We have obtained best tree in Random Forest Regressor at depth=20. Below is the graph which shows features with highest information gain. Feature importance in Random Forest refers to a technique used to understand the significance of input variables (features) in predicting the output variable (target) using a Random Forest model. The Random Forest algorithm provides a straightforward approach for evaluating feature importance, which is based on how much each feature decreases the impurity of the split (often measured using Gini impurity or entropy in classification tasks, and variance reduction in regression).

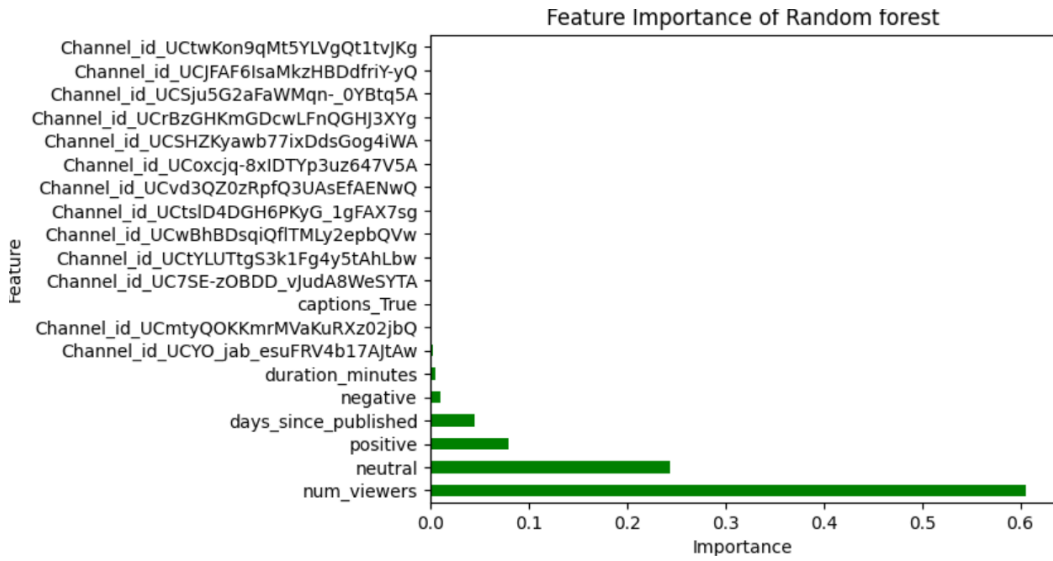


Figure 16: Understanding the Features Influencing Predictions

| | num_viewers | num_fav | duration_minutes | days_since_published | negative | neutral | positive | captions_True | Channel_id_UC4vAON9PsmKoyDqgftx0s-A | Channel_id_UC6-2liodmKiv-ZFNs9VuJ8Mw | ... | Channel_id_UCq46 |
|-----|-------------|---------|------------------|----------------------|----------|---------|----------|---------------|-------------------------------------|--------------------------------------|-----|------------------|
| 42 | 157950 | 0 | 8.433333 | 2564 | 10.0 | 28.0 | 148.0 | 0.0 | 0.0 | 0.0 | ... | ... |
| 215 | 15407 | 0 | 5.100000 | 753 | 6.0 | 9.0 | 20.0 | 0.0 | 0.0 | 0.0 | ... | ... |
| 193 | 389693 | 0 | 11.633333 | 2450 | 14.0 | 31.0 | 101.0 | 0.0 | 0.0 | 0.0 | ... | ... |
| 453 | 89262 | 0 | 10.116667 | 2897 | 7.0 | 10.0 | 24.0 | 0.0 | 0.0 | 0.0 | ... | ... |
| 576 | 17651 | 0 | 20.383333 | 2292 | 5.0 | 3.0 | 21.0 | 0.0 | 0.0 | 0.0 | ... | ... |

Figure 17: Snapshot of data before regression

| Metric | Value |
|--------|-------|
| MAE | 1128 |
| R-Sq | 0.95 |

Below is the snippet of how output table looks like alongwith true values of number of views.

| | Num_likes | Predicted_Num_likes |
|-----|-----------|---------------------|
| 589 | 2271.0 | 5108.956334 |
| 720 | 1081.0 | 1146.438030 |
| 374 | 516.0 | 568.052235 |
| 478 | 401.0 | 442.932694 |
| 574 | 1692.0 | 1878.830750 |

Figure 18: Snapshot of Random Forest predicted values

7.3 LGBM Regression of number of views

In this study, we employed the LightGBM (Light Gradient Boosting Machine) regression model to predict the number of views a video would receive on the YouTube platform. Our predictive model utilized various parameters extracted from the video metadata and user interactions, including the number of comments, presence of captions, number of likes, number of favorites, polarity score of

comments, video duration, and the number of days since the video was published. By leveraging these features, we aimed to understand the factors influencing viewership and provide insights into predicting video performance on YouTube. We have used GridSearch to find best hyperparamater by varying max depth, number of leafs etc., by optimising mean absolute error. We Below is the glimpse of the dataset before and after transformation for this regression analysis. The results of evaluation metrics are tabulated as follows.

| Metric | Value |
|--------|-------|
| MAE | 1333 |
| R-Sq | 0.94 |

| | Num_likes | Predicted_Num_likes |
|-----|-----------|---------------------|
| 589 | 2271.0 | 2811.401852 |
| 720 | 1081.0 | 1196.680802 |
| 374 | 516.0 | 587.699274 |
| 478 | 401.0 | 519.800037 |
| 574 | 1692.0 | 1490.263637 |

Figure 19: Snapshot of LGBM predictions

The table above presents the predicted number of views generated by our LightGBM regression model alongside the actual number of views observed for a sample of five videos from the dataset.

7.4 XGBM Regression of number of views

Here, we have deployed XGBoost (Extreme Gradient Boosting) regression methodology to predict the number of views of YouTube videos. Leveraging a comprehensive array of parameters sourced from video metadata and user engagements, our predictive framework delved into key indicators such as comment count, presence of captions, likes, favorites, sentiment analysis scores, video duration, and publication age. By harnessing the robust capabilities of XGBoost, we endeavored to unearth the nuanced dynamics driving viewership patterns, aiming to provide actionable insights for enhancing video performance and engagement on the YouTube platform. We have used GridSearch to find best hyperparamater by varying max depth, nestimators etc., by optimising mean absolute error.

| Metric | Value |
|--------|-------|
| MAE | 1124 |
| R-Sq | 0.95 |

| | Num_likes | Predicted_Num_likes |
|------------|------------------|----------------------------|
| 589 | 2271.0 | 4322.191895 |
| 720 | 1081.0 | 1157.703979 |
| 374 | 516.0 | 571.750977 |
| 478 | 401.0 | 442.534393 |
| 574 | 1692.0 | 1679.945557 |

Figure 20: Snapshot of XGBoost predicted values

The table above presents the predicted number of views generated by our XGboost regression model alongside the actual number of views observed for a sample of five videos from the dataset.

7.5 Comparison of Model performance

7.5.1 Interpretation of Metrics

Mean Absolute Error (MAE): This metric measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences are weighted equally. A lower MAE value indicates a better fit of the model to the data.

R-squared (R-sq): This is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The higher the R-sq, the better the model fits your data. An R-sq of 0.95 means that 95% of the variance in the dependent variable is predictable from the independent variables.

LightGBM shows a slightly lower R-sq of 0.94, indicating a somewhat less accurate fit to the variance in the data compared to Random Forest and XGBoost. Both Xgboost and Randomforest show very similar performance in terms of R-sq, each achieving a score of 0.95, which is excellent. The MAE is very close as well, with XGBoost performing slightly better (1124.83) compared to Random Forest (1128.83). This suggests that, on average, XGBoost's predictions are closer to the actual values. However, we can prefer Random Forest over XGBoost for its straightforward interpretability regarding feature importance.