Allison Biediger
DSC 680
Project 1-Milestone3

Predicting YouTube Video Performance for Enhanced Viewer Engagement

Business Problem

Youtube is a widely used platform for content creators to grow their followers and earn commission, thus making understanding key factors of video success essential for a channel to thrive. Through this I plan to address the challenges content creators may face when predicting the performance of their videos. To do this I will explore the key factors of the Youtube analytics data including likes, comments and views. By focusing on the video characteristics I will develop a predictive model to give content creators actionable insights to optimize their channels through specific content strategy to help them reach their maximum earning potential.

Background Problem

Youtube has become one of the leading video based platforms offering creators the opportunity to make money through creative outlets to deliver content to viewers. Through the youtube algorithm engagement metrics such as likes, comments and shares are typically used to indicate a video's success. To best manage these metrics creators need data driven insights to help drive strategies for video success, this analysis will give a deeper understanding of factors at play in video success.

Data Explanation

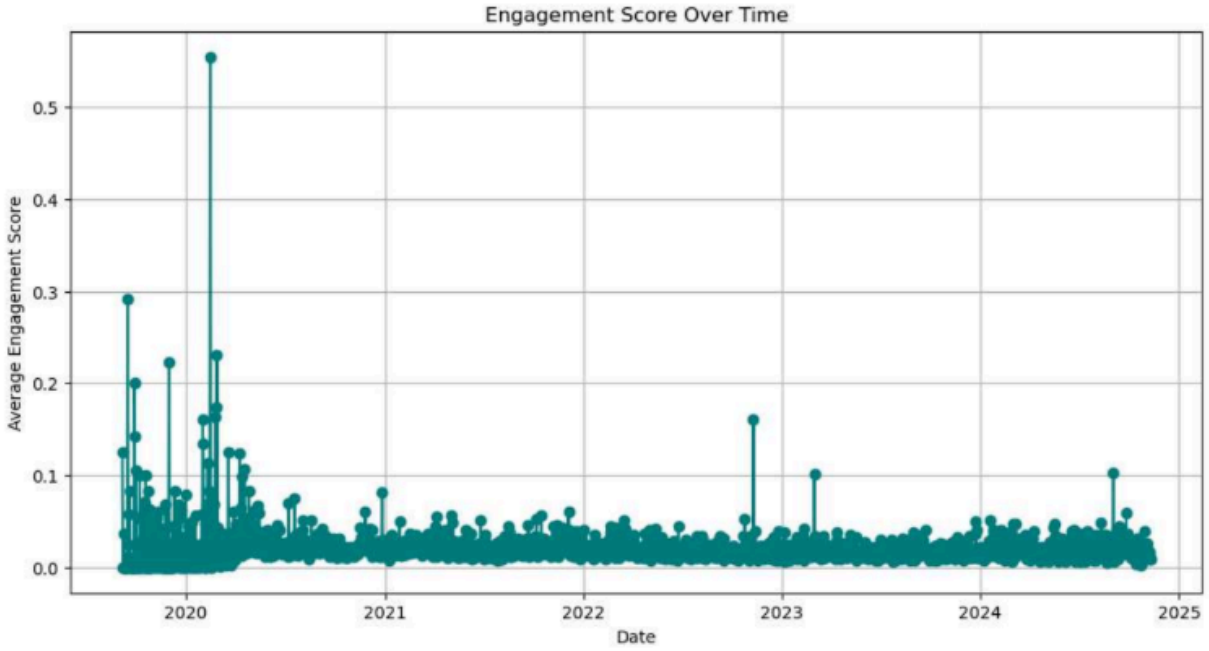
For this analysis I will be using the 200k YouTube Channel Analytics dataset from Kaggle. This dataset contains video metric data including views, likes, comments, shares, and average view duration. To prep the data it was cleaned, duplicates were removed and negative values in likes were removed as well to streamline the analysis. An engagement score was also calculated for each video with the formula {(likes + comments + shares) / Views} to help engage in the analysis.

Methods

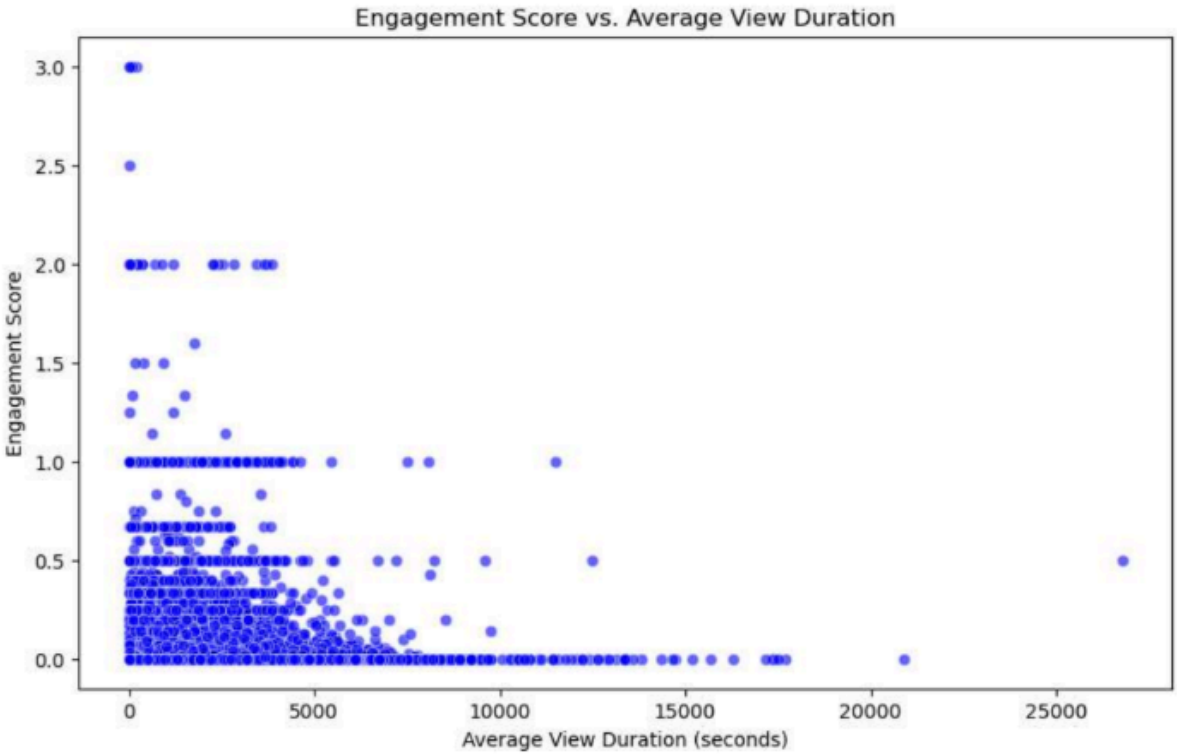
At the start I focused on understanding the data and cleaning it by checking for and handling any missing variables as well as removing negative values from the likes column to further clean the data. From here I focused on doing exploratory data analysis to identify trends, patterns and relationships in the data. I also did some feature engineering to create valuable metrics to gauge the engagement of viewers including managing likes, comments and shares. Then with the new features I developed a random forest model to predict video performance. With the model I evaluated the metrics such as R squared, mean absolute error, and root mean square error to evaluate how successful the model ran. Once I evaluated the model I analyzed the most important factors in the analysis.

Analysis

The analysis of the engagement metrics led to many key findings that can help inform content strategies going forward.



One significant finding was the spike in daily engagement scores around 2020. This is likely due to the increased online presence of people during the COVID-19 pandemic. This heavily showcases the influences of external factors such as major global events on effecting viewer behavior and engagement.



When analyzing the relationship between engagement score and average view duration it was found that shorter videos tended to perform better as far as engagement of the viewers. However when video length watched exceeded 5,000 seconds engagement the video engagement tended to decline. This shows the importance of making concise and impactful videos that will retain the attention of the videos.

In order to dive deeper into the analysis a random forest regression model was made. The performance metrics showed a Mean Squared Error of 0.00037, a root mean squared error of 0.019 and an R squared of 0.9363. These values showed a high predictive accuracy making it a reliable tool to understand the importance of the features at play in engagement. The estimated minutes watched was found as the most important feature with an importance score of 0.5934 reinforcing the importance of this metric. Following this average view duration with a score of 0.1435 and shares with a score of 0.1160 were also important contributors when it comes to engagement. Finally likes with a score of 0.0955 and average view percentage with a score of 0.0256 showed a moderate amount of influence with comments a score of 0.0230 and playlist factors had little to no impact.

Conclusion

This analysis highlights the many factors at play in audience engagement on youtube, showcasing the important factors that are driving the engagement of viewers including watch time, average view duration, and interactions such as likes and shares. The findings showed that shorter videos with high watch times and active participation of the viewer had the highest engagement. There was a spike in engagement during 2020 likely due to the COVID 19

pandemic that highlighted the impact of external factors on youtube engagement. Using a random forest regression model further validated these findings, with the high predictive accuracy and the feature importance rankings also emphasizing the importance of factors such as estimated minutes watched and average view duration. By applying the insights of this analysis content creators can craft data driven strategies that will help optimize engagement, helping to continue to grow an engaged audience.

Assumptions

For this analysis there were many assumptions made to find actionable insights. First I assumed an engagement score formula that would factor in likes, comments and shares while keeping the number of views on the video in mind to better reflect the engagement a video receives. Another assumption is that the dataset reflects a diverse collection of videos through many different genres and creators to give a good variety in the overall analysis. Finally I assumed that external factors and world events (such as COVID 19) influenced behaviors online during the timeframe of the analysis. These assumptions were needed to give a strong foundation to the analysis and potential ways to grow from them.

Limitations

There were many limitations that affected the analysis. One of the major limitations was that on the amount of data, further information regarding the video genres would have given better insights into audience preferences and specific trends. While the dataset did include views and average view duration it did not include the specific information of video length. This would have given a better understanding of how video length may influence engagement metrics. Finally there was no further information on how the videos were promoted, when the video was uploaded and any algorithm requirements or changes which all may have influenced the engagement levels of the videos. These limitations showcase great ways to continue this research with more comprehensive datasets.

Challenges

Several challenges happened during the analysis, one of the most challenging being the interpretation of the nuances of engagement. This metric was influenced by many different factors such as likes, comments, and shares which may not have fully showcased the true outcome of the video. The lack of genre information and video length data added another layer of challenge to find valuable insights. These challenges help showcase the need for a strong understanding of your dataset and the need for a well rounded dataset that contains all needed information.

Future Uses and Additional Applications

This analysis can help guide content creators on youtube to best optimize their strategies to boost their engagement. Future studies can incorporate more variables such as video length and

genre to give a better understanding of engagement drivers in different genres. These same methods can be used in other social media platforms to find cross platform trends and significant factors. Finally this can also help forecast engagement trends to best help creators prepare for audience needs and preferences.

Recommendations

Based on the findings of the analysis, creators should focus on making videos with an average length of 5,000 seconds to maximize viewer engagement. By encouraging likes and comments as well as encouraging sharable content can help further boost engagement scores. Finally creators should also consider exploring genre and video length data in future studies to best optimize strategies.

Implementation Plan

To start implementing these recommendations creators should start by looking through their existing content to find their typical engagement scores and to focus on videos that have had high engagement in the past. This in combination with the recommendations can help the content creator develop a strategy for their content focusing on similar themes or formats. Creators can also leverage analytics tools that will allow them to track some of these metrics in real time. This will allow for a continuation opportunity to optimize their content. Additionally creators can occasionally experiment with different formats or lengths to gain more insights to help evolve their content strategy.

Ethical Assessment

With this data there are many ethical considerations that were kept in mind through the analysis. This includes data integrity,usage of the insights and bias. Through this I noted all changes I make to the data set and all steps in the EDA and modeling process to give repeatability to the process to help ensure the data integrity. There is likely a bias in the data due to algorithm bias that models may showcase as well and the insights in connection may unfairly benefit specific creators. To address these concerns I followed ethical guidelines, transparently managed the data and advocated for fair use of the analysis results.

Questions and Answers:

1) *Why is "estimated minutes watched" the most imporWhy is "estimated minutes watched" the most important feature for predicting engagement?*
    a) This metric reflects the amount of time viewers spend on the content, which directly correlates with their level of engagement. It is combining both viewership and watch time numbers creating a solid indicator of viewer interaction.
2) *How could the lack of video length data have impacted the findings?*
    a) Video length data could have provided better insight into whether shorter or longer videos drive more of a type of engagement. For example if a video is an hour long versus if a video is 5 minutes long viewers may not click on a video that they consider too long or too short.
3) *Are likes, comments, and shares reliable metrics indicating viewer engagement?*
    a) These metrics are commonly accepted to act as indicators of engagement in many social media platforms as they do showcase active viewer participation. However, they may not capture the passive engagement a viewer may have, such as watch time and should be considered alongside other factors like "estimated minutes watched"
4) *What external factors could influence engagement scores?*
    a) Trending topics, world events, social trends and even a potential change in the youtube algorithm set up can cause potentially large impacts on engagement for a video. For example videos speaking about current world events may get a more immediate surge in engagement that shifts going forward.
5) *How were the engagement metrics selected for the engagement score?*
    a) The metrics were chosen based on measurable interactions that reflected different aspects of potential viewer engagement such as likes, comments and shares. By dividing the sum of these by the amount of views it helped standardized the data.
6) *Can this engagement analysis be applied to smaller channels?*
    a) Yes, these methods can be scalable to better suit a smaller channel focus with a combination with adjustments to account for the lower data volumes.
7) *What challenges may be faced when expanding the analysis to emerging platforms?*
    a) Challenges would likely include what data is made available, the difference between used algorithms and user behaviors that are specific to a platform. This would require careful planning.
8) *Are there any ethical considerations that need to be considered for creators?*
    a) Yes, it is important that content creators promote genuine engagement through meaningful content rather than using manipulative practices like clickbait.
9) *How could this analysis assist those just beginning their content creation journey?*
    a) New creators can use these insights to build effective data based strategies for their content creation to encourage engagement. This will help them compete with more solidified content creators in their genre.
10) *What role does watch time play in engagement score?*
    a) Watch time is an important metric that reflects the audience retention and was significant in the factor analysis.