HIGH SCHOOL EDITION
Journal of Student Research

Volume 13 Issue 1 (2024)

An Assessment of YouTube Educational Video Quality Through Machine Learning

John Weng

Andover High School, USA

ABSTRACT

This paper presents a machine learning-based approach to assessing the quality of educational videos on YouTube. With the growing number of educational videos available on the platform, verifying efficacy and credibility has become increasingly important. To address this issue, we propose a framework that allows for automated evaluation of video content based on analytic parameters provided by the YouTube Data API. We look at features such as view count, comment polarity, channel subscriber count, like/dislike ratio, and comment count to capture various dimensions of video quality and engagement. We introduce our target variable as the "Educational Quality Score" (EQS), which measures a video's impact on education using the RACED model - Relevance, Accuracy, Clarity, Engagement, and Depth. Our method offers educators and content creators a tool to enhance their selection of quality educational YouTube videos.

Introduction

In recent years, online platforms such as Coursera, edX, and YouTube have revolutionized the way we consume educational media. YouTube, in particular, offers a diverse range of educational videos spanning nearly every subject. However, as the availability and accessibility of educational videos on YouTube continue to expand, ensuring their quality and effectiveness becomes a critical challenge. Given that YouTube does not scrutinize the credibility of video creators, content that is unsuitable or lacking in expertise frequently finds its way onto the platform (Lee et al., 2022). Furthermore, studies have shown that YouTube may not be as effective for certain educational subjects, such as surface anatomy, where it falls short of delivering comprehensive knowledge (Azer, 2012). As a result, students must exercise caution in relying on the information presented in YouTube videos. This paper focuses on using machine learning techniques to quantify and evaluate the educational value of educational videos. The goal is to create a framework that enables automated evaluation of video content, providing learners, educators, and content creators with a more comprehensive understanding to enhance the learning experience. The key motivation behind this research is due to the dynamic nature of educational videos. Unlike traditional static educational resources, videos can vary widely in terms of content, presentation style, and engagement. Therefore, building a model to analyze the video content for correctness and engagement could prove to be challenging. Hence, this paper provides a novel method for predicting educational quality based on analytic parameters provided by the YouTube Data API.

Background

Video as a form of media has always been a cornerstone of education. It can often make abstract topics easier to visualize and understand (Brame, 2016). Undoubtedly, integrating videos into education can yield a net positive impact. The question is, What are the essential factors that determine the quality of a video for educational purposes? Brame’s research identified three main criteria: appropriate cognitive load, maximization of student engagement, and student interaction. The emphasis on appropriate cognitive load recognizes that videos must strike a delicate balance between presenting essential information and avoiding overwhelming learners. If video content is inappropriate and unsuitable for the audience, the video can be deemed irrelevant. The second criterion emphasizes the importance of maximizing video engagement. Videos are more likely to be more educational if they incorporate interactive elements to maintain students' interest. Finally, the student interaction criterion of Brame's work stresses the value of active learning. A high-quality video is characterized by the active participation of students in its content, such as when they engage in dynamic discussions about the material presented. In the context of YouTube videos, students can actively participate by posting comments. Taking these three criteria into account has been a focus of this project. By building upon Brame's foundational work, this project seeks to contribute to the existing literature on educational video quality assessment and offer insights and methods to enhance the overall quality of online video education.

Dataset

Due to the lack of datasets on YouTube videos specifically in the educational category, an original dataset was created based on existing data provided by the YouTube Data API. By leveraging this API, new videos can be easily added to the dataset if needed. This project utilized Python scripts from a GitHub repository (Weng, 2023) to build the dataset and train and evaluate the models.

Features

The features of this dataset have been selected to capture various dimensions of video quality, engagement, and content attributes. While data reflecting the trends of likes/dislikes on specific videos and engagement metrics like view time could provide valuable insights into video educational quality, such information remains inaccessible for channels not owned personally. Therefore, the following features were decided upon:

View Count: This is a popularity-based metric reflecting the degree of exposure a video has received on the YouTube platform. Incorporating such a metric provides general information on the perceived significance of the video within the educational category.

Comment Polarity: Considering positive or negative user feedback on a video can help us determine whether or not a video is informative and impactful. Comment polarity is determined through a weighted average of polarity values from the top 5 comments in a video. This average is calculated by assigning weights based on the sum of likes and replies for each comment. To automate collecting comment polarity data, this project used Python's TextBlob library to output a polarity value between -1 and 1.

Channel Subscriber Count: A higher subscriber count would underscore a channel's credibility, making it more likely to produce higher-quality content.

Like/Dislike Ratio: A higher like/dislike ratio generally indicates the prevalence of positive sentiment and overall positive viewer feedback, while a lower like/dislike ratio could signify content-related shortcomings with educational objectives.

Comment Count: This feature provides insight into the extent to which a video has sparked curiosity and debate among its viewers. A higher comment count signifies a heightened level of viewer engagement in the content.

Obtaining Initial Data

An initial search consisted of querying the API for key terms that encompassed 10 subjects: economics, engineering, history, literature, mathematics, politics, psychology, science, sociology, and technology. Only videos from the YouTube education category were collected to ensure the relevance of video content. The search term used for each subject is the name of the subject in lowercase as mentioned above. A sample query is provided below:

https://www.googleapis.com/youtube/v3/search?key=[API_KEY]&part=snippet&videoCategoryId=27&type=video&maxResults=50&relevanceLanguage=en&q=[subj_name]&order=relevance

In the query above, subj_name represents the search term that was used for each subject. Fifty videos were taken from each of these individual queries and their corresponding video IDs were extracted and put into a text file.

Assembling Data

To obtain more data, YouTube's "Learning" channel was searched. The channel contains 9 playlists, each focusing on a different educational topic. The video IDs of the top 7 videos from each playlist were extracted. All video IDs were then assembled into one file and the duplicates were removed. The features corresponding to each video ID were added into a CSV file using a similar API querying method as described in the previous section. The final result included 752 unique videos spanning a wide range of educational subjects.

Target Variable

Central to our research is the formulation of a tangible metric that measures the educational quality of YouTube videos. We introduce the "Educational Quality Score" (EQS) as our target variable—a quantified representation of a video's educational value. Research is scarce regarding the topic of an educational quality score (EQS) for YouTube videos. Thus, the scoring criterion used in this paper draws upon related literature in the medical field. A medical paper evaluating the reliability and quality of YouTube videos with "ovarian cyst" content used a scoring criterion that accounted for the accuracy, clarity, and depth of videos using the DISCERN and GQS evaluation models (Andan & Aydin, 2022). It is mentioned in the article that scores range from 1 to 5, where a score of 1 indicates that a video is of poor quality, and 5 indicates that a video is of excellent quality. This study suggests that using factors such as accuracy, while taking into account bias and relevance, and finally assigning a score that is a positive integer between 1 and 5 is feasible. A paper by Kuru and Erken evaluating the quality of YouTube videos as a source of education for rotator cuff tears takes a similar approach to assigning a numerical value to video quality. The article uses the JAMA model for evaluation, consisting of 4 criteria: "Authorship, Attribution, Disclosure, and Currency" (Kuru & Erken, 2020). The options for scoring each criterion are either "Yes" or "No", corresponding to a rating of 1 or 0, respectively. Thus, the maximum score for a single video is 4 with the minimum being 0. Overall, the literature suggests that given the potential variance in video evaluation criteria, using assessment parameters tailored to effectively capture your research objectives ensures an optimal outcome.

Scoring Criteria

The scoring criteria used in this paper is known as the "RACED" model, an acronym for Relevance, Accuracy, Clarity, Engagement, and Depth. Below is a short description of each criterion.

Relevance: The "Relevance" dimension measures the alignment of content with educational subjects, ensuring that videos contribute positively to the viewers' learning.

Accuracy: The "Accuracy" dimension measures the factual correctness and reliability of video content, ensuring that the information presented aligns with educational objectives.

Clarity: The "Clarity" dimension gauges the video's capacity to convey concepts with clarity, measuring the degree to which it creates an optimal learning experience for viewers.

Engagement: The "Engagement" dimension evaluates the degree to which videos foster active viewer participation, as evidenced by comments, likes, dislikes, and the extent of shared discussions.

Depth: The "Depth" dimension measures the extent to which the video delves into the complexities of the subject matter, evaluating a video's potential to provide a comprehensive understanding of the video topic.

All criteria are scored as an integer ranging from 0 to 2, inclusive. A score of 0 indicates that the video did not meet the expectations for a particular criterion. A score of 1 indicates that the video somewhat met the expectations for a particular criterion. A score of 2 indicates that the video met all the expectations for a particular criterion. These 5 criteria total 10 points, implying a minimum score of 0 and a maximum score of 10. The scores were manually assigned for each video.

Methods

Data Preprocessing

Some of the gathered videos exhibited characteristics that did not align with the essential components of one or multiple target variables. To clean the data, videos were deleted from the original dataset to ensure optimal results. Videos were removed if:

- They had an undefined like/dislike ratio, i.e., the number of dislikes was 0.
- The creator of the video did not enable likes or dislikes.
- They had less than 5 total comments.
- No channel subscriber count was available from the data gathered.
- The video content and/or transcript was not in English.
- The video was not intended to be educational, e.g., advertisements, entertainment videos, comedy, etc.

In total, 250 videos were removed from the original count, leaving 502 videos for the model. In addition to preprocessing, an 80/20 training to testing data split was used.

Model

To predict a continuous value between 0 and 10 for EQS is a regression problem. We used 3 popular regression models: RandomForest (Pedregosa et al., 2011), CatBoost, and XGBoost. The rationale behind selecting Random Forest was to gain insight into which features were important in determining video EQS.

Hyperparameters

To optimize each of the models, we performed hyperparameter tuning. The hyperparameters for our models are listed below:

RandomForest:

- n_estimators = 150

XGBoost:

- Objective = reg:squarederror
- n_estimators = 100
- Learning_rate = 0.15

- Max_depth = 4

CatBoost:

- Iterations = 1000
- Depth = 6
- Learning_rate = 0.1
- Loss_function = RMSE

Results

After running the models on the dataset, the following results were obtained. Table 1 below shows the performance of each of the models, measuring their R-squared, mean squared error (MSE), and mean absolute error (MAE) values. In the table below, MSE and MAE are rounded to the nearest hundredth.

Table 1. Model Statistics on Training and Testing Data

Model	Training			Testing		
	R^2	MSE	MAE	R^2	MSE	MAE
RandomForest	0.93	0.10	0.23	0.56	0.58	0.58
XGBoost	0.95	0.07	0.20	0.61	0.51	0.56
CatBoost	0.99	0.01	0.03	0.54	0.56	0.55

Model Statistics and Performance

The following graphs show the performances of each of the models on the testing data consisting of 101 rows. Corresponding to each model is a graph of the predicted versus actual EQS values. A graph with more data points deviating from the diagonal red line indicates less accuracy. A darker shade of blue corresponds with a greater frequency of a data point whereas a lighter shade indicates less frequency.

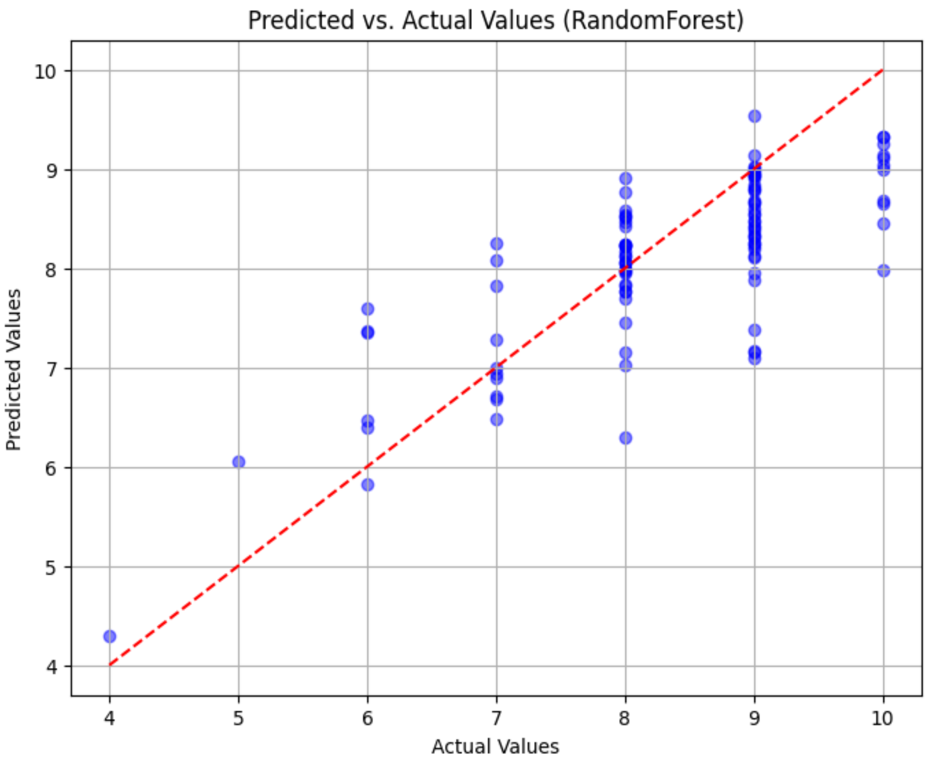

Figure 1. Predicted Vs. Actual Values on Training Data (Random Forest)

Random Forest, with the highest MSE (0.58), shows high variance when scores are between 7 and 8, inclusive, and when scores are equal to 10. It also displays a consistent negative bias in predicting EQS scores 9 or greater, indicative of an overall underprediction.

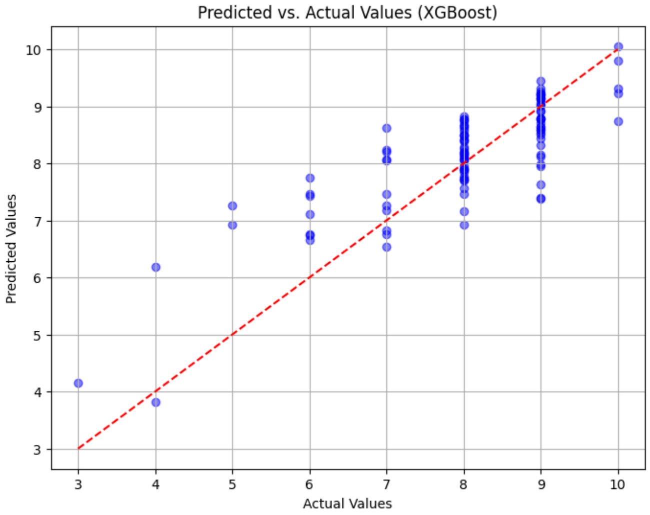

Figure 2. Predicted Vs. Actual Values on Training Data (XGBoost)

XGBoost exhibited the lowest MSE (0.51) on the testing data among the three models, indicating best overall performance. However, the model also showed a slight negative bias in predicting scores 9 or greater with a positive bias in predicting scores between 3 and 8, inclusive. Hence, XGBoost overpredicts scores overall.

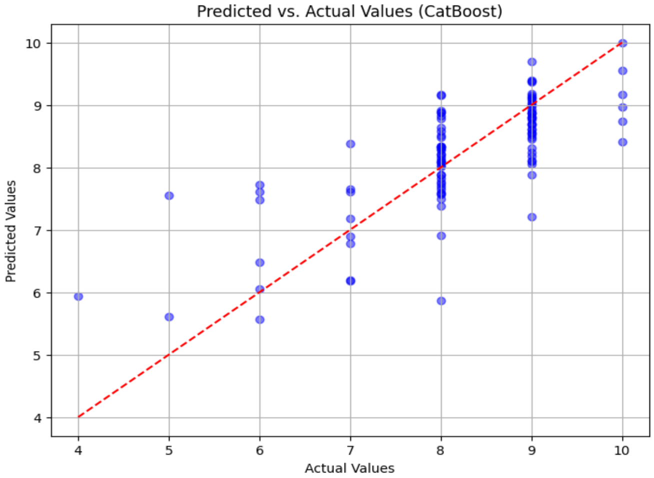

Figure 3. Predicted Vs. Actual Values on Training Data (CatBoost)

CatBoost, while also accurate (MSE = 0.56), shows the greatest variance in scores less than or equal to 7. Similar to the other models, CatBoost exhibits a negative bias when the EQS score is 9 or greater, indicating an overall underprediction.

Additionally, histograms of residuals are provided, where each residual is calculated as y_{test} - $pred$. y_{test} is the actual value of the target variable in the dataset, and $pred$ is the value predicted by the model. If a distribution is skewed towards the left, $y_{\text{test}} - pred < 0$, meaning $pred > y_{\text{test}}$, implying that the model is overpredicting the target. If a distribution is skewed towards the right, $y_{\text{test}} - pred > 0$, meaning $pred < y_{\text{test}}$, implying that the model is underpredicting the target. In a histogram of residuals, a normal distribution indicates that the model's predictions are unbiased, on average, and that the errors are due to randomness, suggesting that the model captures the relationships in the data well.

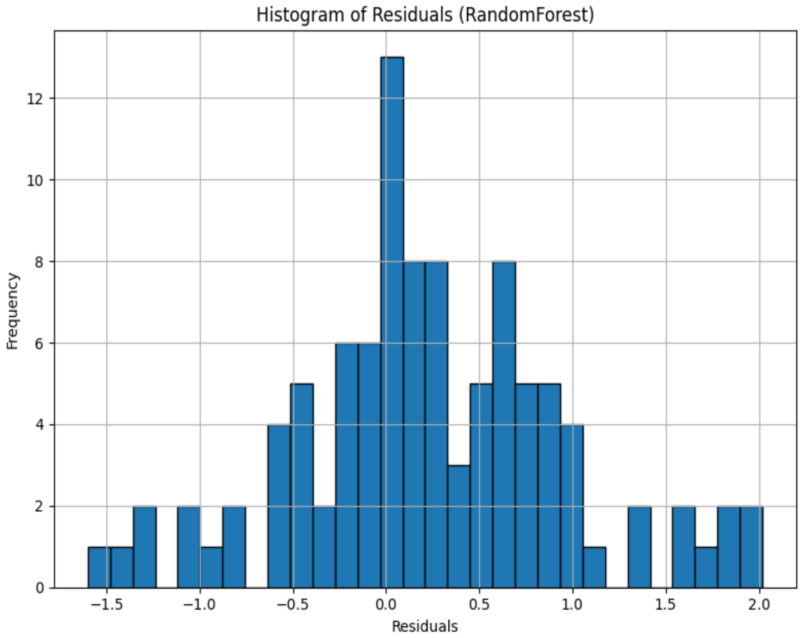

Figure 4. Residual Histogram (RandomForest)

RandomForest's residual histogram exhibits a right-skewed distribution, where the majority of values are to the right of 0. This suggests that on average, RandomForest tends to underpredict scores, which is also evident from the scatter plot of predicted vs. actual values.

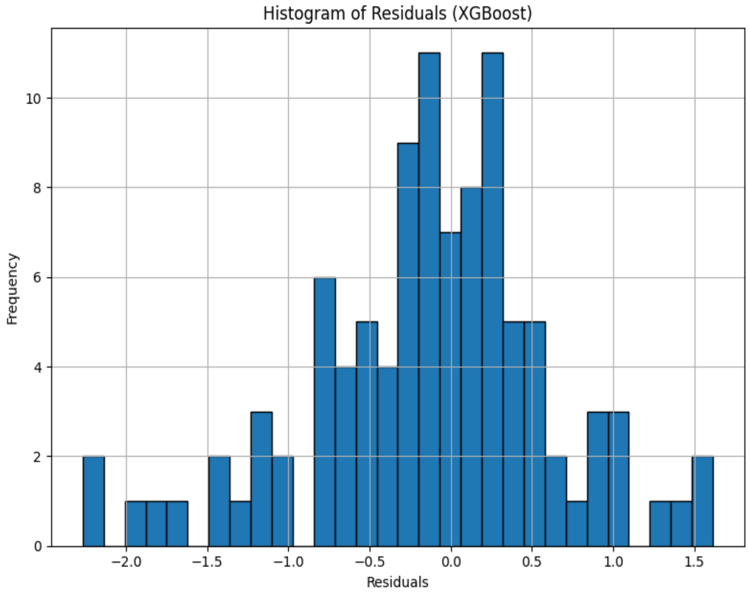

Figure 5. Residual Histogram (XGBoost)

The residual histogram for XGBoost shows a left skew. There is a higher frequency of data that is extended towards the left of 0. This indicates that on average, XGBoost tends to overpredict EQS scores (also supported by the scatter plot).

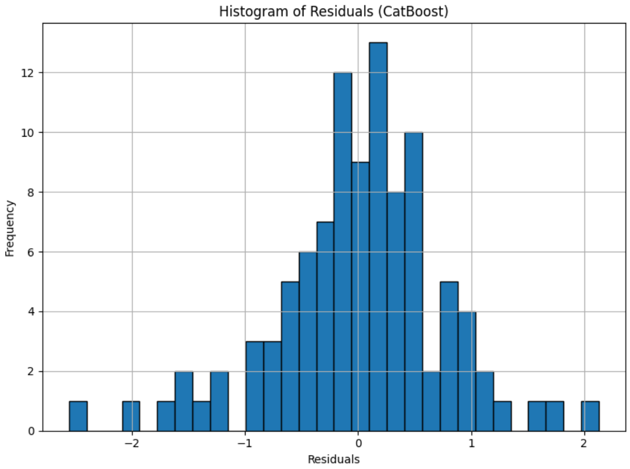

Figure 6. Residual Histogram (CatBoost)

The distribution of residuals in the histogram above has a slight rightward skew, with more values clustered to the right of 0. As indicated by the predicted vs. actual scatter plot, CatBoost predicted scores are lower than actual scores where the values are 9 or greater. This accounts for the overall underprediction.

A feature importance plot shows the significance of each feature in predicting the target variable in a dataset. Features with higher scores have greater influence on a model's predictions. Below is a feature importance plot generated from our Random Forest Regressor model.

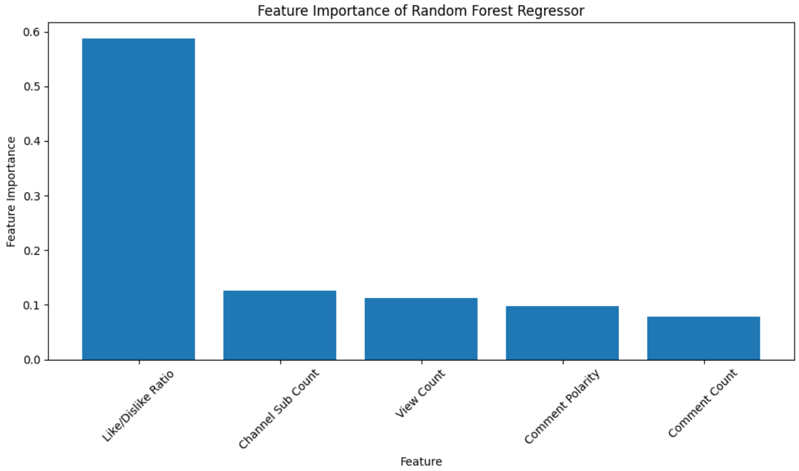

Figure 7. Feature Importance (RandomForest). Feature weight values from left to right (rounded to the nearest hundredth): 0.59, 0.13, 0.11, 0.10, 0.08

Analysis of Results

From the data in Table 1, we see that RandomForest had the worst performance considering squared error and absolute error metrics. This is most likely because gradient-boosting algorithms such as XGBoost and CatBoost are, on average, better in terms of performance than RandomForest (Gupta, 2021). Overall, XGBoost seems to demonstrate the most favorable statistics. One reason for this could be that the initial dataset had a notable imbalance, with the majority of videos having an EQS score ranging from 7 to 9. Gradient boosting algorithms such as XGBoost and CatBoost give more weight to certain variables after failing to predict them initially, making them better for more unbalanced datasets, whereas it is not guaranteed that random forest algorithms will do the same. Additionally, the MSE and MAE values are higher for the training data than for the testing data, indicating an overfitting issue. Perhaps in future projects, this problem can be mitigated by increasing the amount of training data.

From the feature importance data shown in Figure 7 above, we see that like/dislike ratio has the highest importance by a large margin, indicating that like/dislike ratio is the most important in predicting overall video quality. The reason for this is because the like/dislike ratio is an overall feedback metric that takes into account both positive and negative feedback. On the other hand, while channel subscriber count, view count, and comment count serve as effective metrics for assessing popularity and providing a broad understanding of video engagement, they do not explicitly incorporate user feedback. Another explanation for why channel subscriber count is not as important for predicting video quality could be that the size of a channel does not necessarily reflect its ability to produce videos of superior quality. In addition, it is rather unexpected that the sentiment polarity of comments holds minimal significance in determining the overall quality of a video. A possible explanation is that comments can be diverse and complex, and could involve content that is too complicated for text processing libraries to understand. Another factor to consider is that the sentiment polarity ratings may register as negative on a video discussing a negative subject matter. However, this does not necessarily reflect the actual quality of the video content itself.

Conclusion

In this project, we have successfully built models to predict an EQS for YouTube educational videos. These models can then be used by instructors and learners alike to find out the quality of videos. Educators can save time by avoiding low-quality videos. Over time, as more data is added to the dataset, these models can show factors that contribute to the quality of YouTube educational content. Of course, this project is not without limitations. Some future directions to improve model accuracy could be considered, such as more advanced natural language processing (NLP) techniques to extract video content accuracy and topic relevance. In addition, the video transcript could be another feature that could be analyzed to identify educational terminology, coherence, and depth of information. Furthermore, video and/or comment credibility could be considered to yield a more accurate result. In conclusion, this project shows the potential of machine learning to assist users in selecting high-quality educational content from the vast array of YouTube videos. By developing these models, we have taken a significant step towards providing users with a tool that helps them make informed decisions on video selection in their education. As the field of machine learning and educational technology continues to evolve, our work offers a foundation for future research and innovation in this dynamic domain.

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Ganesh Mani for his guidance and expertise throughout this research project. I also would like to thank my teaching assistant, Ms. Joanna Gilberti for her advice and feedback. Additionally, I would like to acknowledge Scholar Launch for providing me with this research opportunity. Last but not least, I would like to express my appreciation to my family for their support and encouragement during this journey. Thank you all for your contributions and support.

References

Andan C, Aydin M F (March 01, 2022) Evaluation of the Reliability and Quality of YouTube Videos on Ovarian Cysts. Cureus 14(3): e22739. https://doi.org:10.7759/cureus.22739

Azer, S.A. (2012) Can "YouTube" help students in learning surface anatomy?. Surg Radiol Anat 34, 465–468. https://doi.org/10.1007/s00276-012-0935-x

Brame C. J. (2016). Effective Educational Videos: Principles and Guidelines for Maximizing Student Learning from Video Content. CBE life sciences education, 15(4), es6. https://doi.org/10.1187/cbe.16-03-0125

Gupta, A. (2021, April 26). XGBoost versus Random Forest. https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30

Kuru, T., & Erken, H. Y. (2020). Evaluation of the Quality and Reliability of YouTube Videos on Rotator Cuff Tears. Cureus, 12(2), e6852. https://doi.org/10.7759/cureus.6852

Lee, K. N., Tak, H. J., Park, S. Y., Park, S. T., & Park, S. H. (2022). YouTube as a source of information and education on endometriosis. Medicine, 101(38), e30639. https://doi.org/10.1097/MD.0000000000030639

R, S. E. (2021, June 17). Understand Random Forest Algorithms With Examples. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830 https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Weng, J. (2023). EduMLProject. GitHub Repository. https://github.com/jweng2190/EduMLProject