MACHINE LEARNING-BASED REVENUE PREDICTION FROM YOUTUBE CHANNEL ANALYTICS

[1]Dr. V Indumathi
[1]Assistant Professor,
[1]School of Computer Studies- UG- BCA, RVS College of Arts and Science, Sulur
indhumathi@rvsgroup.com

Abstract: YouTube has become one of the most powerful media in terms of creating contents, advertisements, and the creator economy; hence it becomes extremely important to create a proper revenue forecasting as an important part of digital media studies. In this paper, we present the model predicting revenue for a YouTube channel. It was done via performing exploratory data analysis on a dataset consisting of 364 observations with 70 features that measure analytics related to monetization, engagement, temporal, and audience behavior features. In our analysis, we perform data visualization, extract time-based information from the features, explore the data, find correlations between features, and use random forest regression models to predict the revenue. It appears that revenue estimation can be highly correlated with such features like Watch Page Ads Revenue, YouTube Ads Revenue, Estimated AdSense Revenue, Monetized Playbacks, and Ad Impressions, and the result of training random forest on that dataset gives us the value of R^2 equal to 0.99 on hold-out.

KEYWORD

YouTube analytics, revenue prediction, creator economy, Random Forest, machine learning, digital platform analytics.

1. INTRODUCTION

The role of YouTube concerning the posting of videos, interactions with the audience, and making money on the content has significantly increased, and it makes money via advertising, premium services, membership plans, and other user engagement [1][2].Industry analysis and data obtained from different analytic platforms within the period 2024-2026 have confirmed that there are indeed creator-oriented analytics and monetization [4][5] as depicted in Fig. 1.

Figure: 1 Comparing YouTube with Other Platform

It is clear from the industry analysis that the use of quantitative models would help analyze and predict the income flow [3][6]. It has been proved from practical experience that the correct data set was used to conduct YouTube analytics and includes 364 observations with 70 attributes like views, watch hours, subscribers, impressions, estimated earnings, thumbnail click-through rate, monetized views, adSense earning, YouTube ads earning, and YouTube Premium earning.

2. RELATED WORK

According to modern research, it can be stated that YouTube analytics is considered to be performance assessment that considers discovery, retention, engagement, and monetization on an equal basis [3][4][6]. It was stated in YouTube analytics rules that calculations related to the estimation of ad revenue, YouTube Premium revenue, and other monetization factors are independent, proving that creators' income appears due to different revenue sources [1][2]. In addition, the guidelines for practical YouTube analytics in the period of 2022-2026 state that RPM, impressions, CTR, watch time, and engagement are believed to be considered jointly because these indicators impact audience size and ability to earn [3][4]. On the other hand, it should be mentioned that the use of machine learning techniques for forecasting YouTube metrics implies the choice of the appropriate target variables. For instance, in 2021, it was necessary to determine which metrics (e.g., views, subscriptions, watch time, etc.) can predict daily earnings using an ML toolkit to estimate YouTube metrics. As the results showed, tree-based regression was more efficient than other models [7]. In addition, the technique of trend analysis implies the implementation of machine learning algorithms aimed at analyzing various factors related to metadata and interactions with videos, including their category and popularity [8][9]. A [14] paper on an advanced analytics dashboard further emphasized the demand for creator-facing systems that combine descriptive metrics with predictive intelligence to guide channel strategy[10].

In addition, nowadays, there are some free-of-charge sources of information that can be used for developing models of performance for the purposes of analyzing not only different channels but also creators themselves. Thus, for example, income data related to YouTube creators can be found at Kaggle while platforms like Social Blade and vidIQ are examples of those containing information about different metrics that can be taken into account in the decision-making process [11] [5].

3. MATERIALS AND METHODS

3.1. Dataset Description

There are 364 observations with 70 variables in this dataset. In the data frame that is to be presented here, there are some significant columns like video duration, upload date of the video, number of days passed after the video was uploaded, day, month, year, day of the week, earning per 1000 views ($), monetized plays, cost per thousand ($) based on play, cost per thousand ($) based on ad, ad view, earning from ad sense ($), YouTube ads earning ($), watch page ads earning ($), views, watch hours (hours), subscribers, earning estimate ($) impression and video thumbnail click-through rate (%). As observed from the data frame above, there are no NULL entries in all the rows; that is 364.

Table 1. Dataset Summary

Field Description	**Value**
Number of records	364
Number of columns	70
Data types	63 float, 5 int, 2 object
Missing-value status in displayed info() output	No null loss shown
Target variable	Estimated Revenue (USD)
Model used	Random Forest Regressor
Train-test split	80:20
Random state	42
Number of trees	100

Despite the limited number of entries within the dataset, it is rich in terms of its width because all the monetization, engagement, and time variables are represented. By considering the results of using the info() method, it can be stated that all the columns include 364 non-null entries. Thus, we will have no significant problems when preparing for regression in preprocessing. It can be concluded from the indicators that are included in variable design; there are indicators of monetization, engagement, viewership, time, and interactions with respect to the platform.

3.2. Preprocessing and Feature Engineering

Video Publish Time was transformed to the data type of datetime. Then, we extracted the year and month from the date on which this video was published. This step might be quite simple but rather essential, since the age and the publication date might affect revenue generation and the user engagement rate. We also used the selected numeric variables to create the numeric dataframe. In the case of predictions, our target variable is Estimated Revenue (USD). We generated the features' matrix by excluding the target variable and ID columns. It is a logical choice for supervised learning, as it doesn't let you predict the target value based on any unique ID or the target variable itself. After that, we made an 80:20 split for training and testing with random_state = 42. Then, we fitted a model of Random Forest Regressor containing 100 trees.

3.3. Exploratory and Statistical Analysis

There were two types of exploratory analysis that were conducted prior to the model building process. One of these included the creation of the scatter plot between Views and Estimated Revenue (USD). The chart can be seen in Fig. 2. It is called Impact of Views on Revenue and was created with Views placed on the x-axis and Estimated Revenue (USD) on the y-axis. What makes this chart significant is the ability to analyze visually if high traffic leads to positive monetization outcomes. As mentioned before, concerning YouTube metrics, views do not determine revenue generation directly, although it remains a very important factor [3][4].

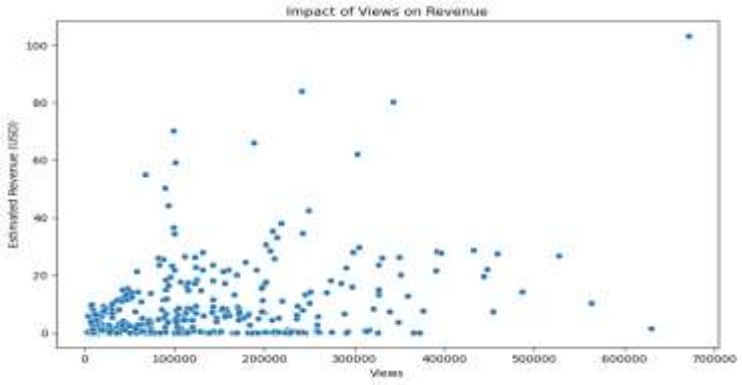

Figure: 2 Views vs Revenue

The next step involved the creation of a bar chart for the mean of Averaged Revenue (USD) by Day of Week as shown in Fig.3. This chart is shown in Fig. 3 below and called Average Revenue by Day of Week. It is utilized in order to compare the average estimated revenue per days of the week following an ordinal categorical variable (from Monday to Sunday).

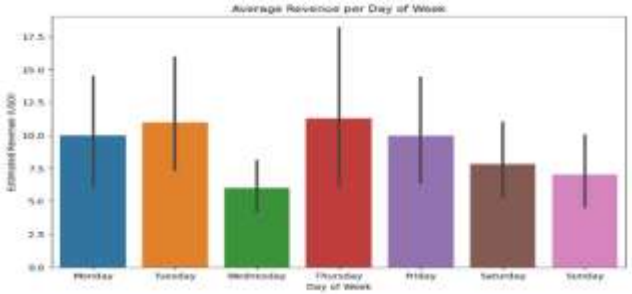

Figure: 3 Average Revenue

This pair of visualizations is to identify, respectively, the general relationship between traffic and monetization and the possible influence of time series on monetization.Calculations of the Pearson correlation coefficients, as shown in Fig. 4, have been done between Averaged Revenue (USD) and all other numeric variables.

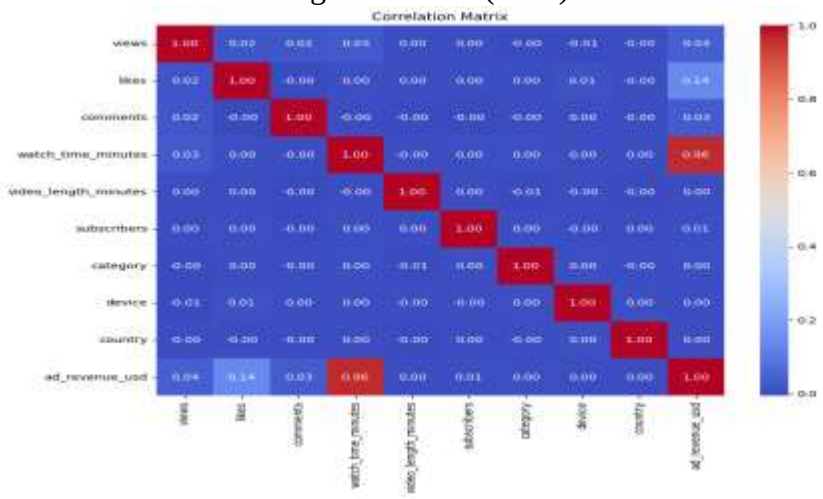

Figure: 4 Correlation Matrix

These were used to generate an ordered ranking of the most strongly correlated revenue-predicting features, and thus served as an informative diagnostic procedure prior to using the non-linear Random Forest model on the data. The choice of the latter model means that, unlike the second-stage regression model, correlations do not play a role in determining the algorithm performance.

Table 1. Top Correlations with Averaged Revenue (USD)

Features	**Correlation**
Watch Page Ads Revenue (USD)	0.999493
YouTube Ads Revenue (USD)	0.999471
Estimated AdSense Revenue (USD)	0.995314
Monetized Playbacks (Estimate)	0.944155
Ad Impressions	0.825534
YouTube Premium (USD)	0.718794
YouTube Premium Watch Time (hours)	0.588183
YouTube Premium Views	0.574533
Clip Views	0.472504

The high correlations seen in Table 1 confirm that the estimated revenue feature is overwhelmingly associated with metrics capturing direct monetization results, in particular those measuring ad-revenue. This finding makes sense because of the economic dependence of the target variable from monetization results. The strong correlation with AdSense and watch page ad metrics is also understandable, considering how monetization metrics are defined in the official YouTube statistics - estimated ad revenue and premium revenue, as opposed to secondary engagement features [1][2]. On the other hand, the high, though somewhat weaker, correlations for Ad Impressions, Premium views, and premium watch time imply that the revenue result is also dependent on exposure quantity and monetization quality factors.

4. Evaluation Metric

The model was evaluated using the coefficient of determination, R^2, on the test split. This metric measures the proportion of variance in the dependent variable explained by the model and is common in regression studies. A value close to 1.0 suggests that the fitted model captures most of the revenue variability present in the test data.The predictive accuracy of the model was measured using the R2 score of 0.99 for the Random Forest Regressor algorithm. The exceedingly high score implies that the algorithm explains virtually all the variance of revenue in the test set of data. While the score is impressive, it must be viewed with a degree of skepticism. Practically speaking, then, the model reached perfect fitting on the given split sample. This validate the result,

but it suggests that future versions of the study should distinguish between pure explanatory prediction and operational forecasting under reduced-information conditions.This follows both current analytics practices and recommendations, where all of the factors are interrelated [3][6][12].. In general, the metrics show that YouTube revenue has a complicated relationship with view count, yet the monetization process remains a key factor.

5. Discussion

The current research offers proof regarding the use of machine learning in predicting the YouTube content revenues by means of analytics. It would seem that the Random Forest regressor algorithm is able to estimate non-linear dependencies between ad revenue, premium revenue, impression statistics, and audience metrics. As a result, predictions become incredibly accurate. In comparison with the simple model, which describes the processes, the current algorithm provides an additional value for a compact and efficient way to predict YouTube content revenue. On the other hand, the coefficient of determination, being extremely high, causes several concerns regarding the methodology. The reason lies in the fact that the key variables proved to correlate with the outcome variables, and in many cases, these variables represent the components of the monetization process. To create a more effective criterion in the future, the comparison of two cases will be conducted, namely the comparison of the cases that include the use of both post-publication and pre-publication variables. The following limitation is associated with the nature of splitting used. Overall, train-test could be considered a random split, which could distort results when dealing with sensitive creator information because of the similarity between observations. Despite modern analytics tools and dashboard approaches that emphasize the possibility of taking actions and predicting outcomes, practical approaches imply stability in time and different monetization conditions [10][13].

6. Conclusion and Future Work

In this research paper, we looked at the studies undertaken through the use of a validated dataset acquired from the Kaggle website in YouTube Analytics. We analyzed a total of 364 records with 70 features through visualization of correlations on revenues. We trained the Random Forest Regressor algorithm which performed satisfactorily with a score of 0.997948 on test split. Top five revenues included Watch Page Ads Revenue, YouTube Ads Revenue, Estimated AdSense Revenue, Monetized Playbacks and Ad Impressions thus implying that the creators' revenue from this dataset is highly affected by the monetization components provided by the platform. As far as future work is concerned, some issues in this study need to be sorted out. Firstly, interpretability of results can be enhanced using SHAP, permutation importance and partial dependence plot methods to ensure that the results are easily interpreted by the researchers and even channel owners. Secondly, time series cross-validation technique could be used instead of simple randomization in testing model performance. Thirdly, benchmarking is required with smaller numbers of feature sets to prevent any revenue leakages into predictions. Finally, future datasets must contain metadata, thumbnails, category tags, title language, retention charts, and text semantics to allow for multivariate analysis of revenue prediction [8][9].

REFERENCES

[1] Metrics | YouTube Analytics and Reporting APIs | Google for Developers. (n.d.). Google for Developers. https://developers.google.com/youtube/analytics/metrics

[2] Understand ad revenue analytics - YouTube Help. (n.d.). https://support.google.com/youtube/answer/9314357

[3] Govorkov, K. (2026, February 5). The Ultimate Guide to YouTube Analytics: 15 Metrics That Matter. Improvado. https://improvado.io/blog/youtube-analytics-guide

[4] Videoboosters, &Videoboosters. (2024, July 2). Exclusive guide to YouTube Analytics: Important metrics for 2024. Video Boosters Club. https://videoboosters.club/2024/05/25/youtube-analytics

[5] YouTube Stats & Channel Analytics - Earnings & Growth | VIDIQ. (n.d.). vidIQ for YouTube.

[6] YouTube Analytics: Key Metrics to Grow Your channel in 2024. (n.d.). https://maekersuite.com/blog/youtube-analytics

[7] Pillai, N. M., L, B. K., B, H. G. S., & R, J. S. (2023). Predictive Modeling of YouTube Using Supervised Machine Learning Algorithm for Identifying Trending Videos and its Impact on Engagement. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.10370614

[8] Meshram, V., Gaikwad, V., Pathak, V., Mohite, A., Barwal, R., & Computer Science Department, SavitribaiPhule Pune University, GenbaSopanraoMoze College of Engineering, Balewadi, Pune, Maharashtra, India. (2023). YouTube trending videos' prediction & analysis. In International Journal of Advanced Research in Computer and Communication Engineering (Vol. 12, Issue 4) [Journal-article]. https://doi.org/10.17148/IJARCCE.2023.1242

[9] Patel, D., 1, Modi, D., 1, & Patel, N., 1. (2022). Trend analysis and prediction of YouTube Videos using Machine Learning Techniques [Journal-article]. Technix International Journal for Engineering Research (TIJER), 9(9), 37–39. https://tijer.org/tijer/papers/TIJER2209005.pdf

[10] Singh, R. P., Kumar, A., Niharika, N., Jain, G., Rastogi, R., & Jain, M. (2024). Development of a Comprehensive YouTube Analytics Dashboard for Enhanced Performance Insights. IEEE Xplorer, 1089–1094. https://doi.org/10.1109/globalaisummit62156.2024.10947784

[11] Kaggel Dataset 2018-2021. (2021). https://www.kaggle.com/datasets/kristhecoder/youtube-revenue-data-20182021

[12] Statista. (2026, March 26). YouTube users worldwide 2020-2029. https://www.statista.com/forecasts/1144088/youtube-users-in-the-world/?srsltid=AfmBOoon4hINyNJme-NdH8l94kxIEJ-xmPBpYaAgU5Eczysvr2HHynvp

[13] Singh, A., Rai, A. N., Saxena, A., Gupta, D., &Bhatnagar, P. (2020). YOU TUBE DATA ANALYSIS USING HADOOP. In International Journal of Creative Research Thoughts (IJCRT), International Journal of Creative Research Thoughts (IJCRT) (Vol. 8, Issue 4, p. 1889) [Journal-article]. https://ijcrt.org/papers/IJCRT2004252.pdf

[14] Cezim, B. (2026). YouTube analytics guide 2025: How to read, track, and grow your channel. Sociality.io Blog. https://sociality.io/blog/youtube-analytics/