Capstone Project Final Report:

‘Exploring the factors that influence YouTube video views through feature analysis and prediction’

By Oyindamola Omole
26.06.23

6,660,986 views

73,556 12,307

Problem Statement and Background

This report examines the key factors that significantly impact YouTube video views. The objective of this project is to offer content creators valuable insights into optimising their YouTube videos to maximise viewership.

Gaining understanding about how to increase both exposure and reach provides a multitude of benefits that are not limited to personal satisfaction and motivation. In society, social media platforms offer opportunities for individuals to generate income through avenues such as monetization and more.

This report presents how Machine Learning Regression models can be used to offer valuable insights into which features are the most significant predictors of the number of views a video obtains. This study narrows its focus on the features that are available upon publishing in order to create predictions that are applicable in real life scenarios. Data that is only made available after a video has gained popularity is not applicable to the project aim.

This study leveraged data from Kaggle's 'Trending YouTube dataset statistics,' which was obtained through web scraping directly from the YouTube website. Regression models were employed throughout the project to facilitate the exploration of the relationship between the continuous numerical variable, views, and its associations with various feature variables.

Data Collection & Processing

The data came in two formats: JSON files and CSV files. CSV files were mainly used during this stage of the project. Cleaning text data was a major part of the data processing stage. The data, obtained through web scraping the YouTube website contained irrelevant information such as numerical values and hyperlinks. These features had to be removed so that the data only provided meaningful values for analysis.

The text ‘Tag’ column was vectorized using CountVectorizer, and the remaining text data was transformed through a column transformer during the modelling stage to prevent data leakage.

EDA

The primary objective of the exploratory data analysis (EDA) was to gain insights into the data distribution and understand variable relationships. This was done through visualisations of box plots, bar graphs, and scatter plots.

During the initial exploration of the target variable, it became clear that the dataset contained a significant number of outlier values. This observation in figure 1 was to be expected given the nature of the dataset, which consisted of ‘trending videos’. Notably, even within the realm of outliers, variations in performance were observed among the videos.
Figure 2 presents the data after adding Log transformation on the data for normalisation.

Figure 1:

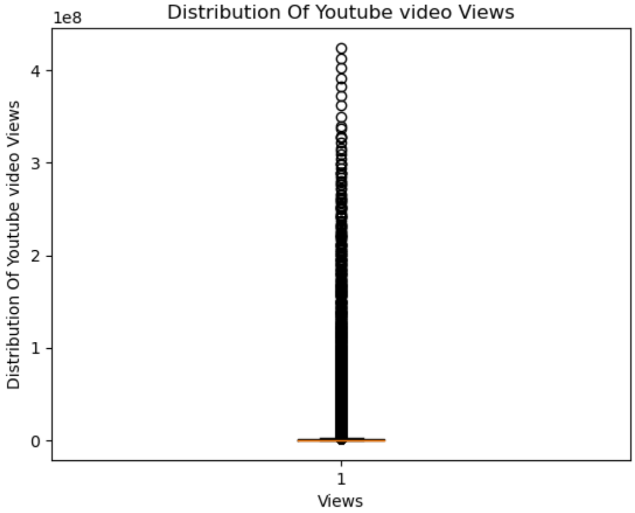
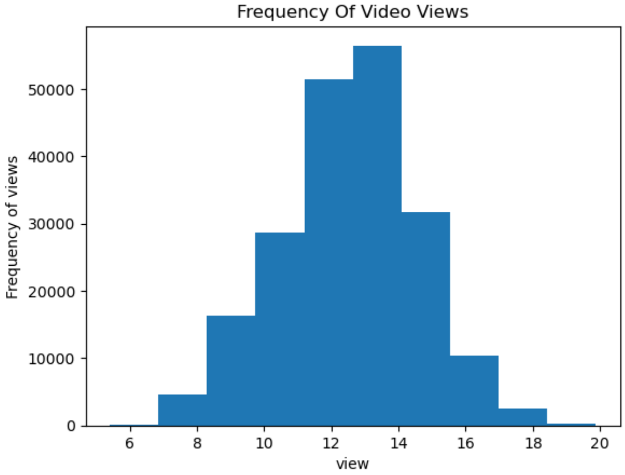


Figure 2:



Data Modelling

Various models were applied and evaluated using metrics such as RMSE, MSE, and R2. Feature selection was performed using Grid Search, Cross Validation, and Pipeline techniques to streamline model fitting, tuning, evaluation, and data preprocessing.

Initially, a baseline Lasso regression model was fitted with one text feature, and subsequently, a second model was fitted with additional features. This approach was used to mitigate data leakage and make the model evaluation reliable. The process also ensured consistent text data vectorization throughout this section.

Modelling Results:

The Baseline Linear Regression Model with Tags only:

- Mean Squared Error (MSE): 0.7283084711045227
- Root Mean Squared Error (RMSE): 0.8534099080187215
- R-squared (R2): 0.8285503318387402

The Baseline model had high MSE and RMSE values, showing bad accuracy in its predictions. Additionally, the R squared score of 0.8285503318387402indicated that the model explained a large portion of the variance in the target variable.

Baseline Lasso Regression Model with Tags only:

- Mean Squared Error (MSE): 1.277964599074526
- Root Mean Squared Error (RMSE): 1.1304709633929242
- R-squared (R2): 0.6991568612391986

The Baseline Lasso Regression model needs improvement in accuracy and explaining the variance in the target variable. The MSE/RMSE values reveal that there were some prediction errors. The R-squared score reveals that the model explained a portion of the variance in the target variable. At this stage adding additional columns to the data was considered.

Gridsearch for Lasso and Ridge Regression model: Added ‘Description’ column:

- Mean Squared Error is: 3.395321857328653
- Root Mean Squared Error is: 1.8426399152652297
- R-Squared Score is: 0.20071394356179806

Chosen model: 'model': Lasso(alpha=0.01), 'model__alpha': 0.01, 'scaling': MaxAbsScaler()

The grid search model's performance was quite poor. The MSE/ RMSE values show a high amount of prediction error. The low R-squared score suggests that the model did not accurately explain the variance in the target variable ‘views’.

Gridsearch for Lasso and Ridge Regression model: Added ‘Description’ column: - Dropping data from post publishing:

Chosen model: 'model': Lasso(alpha=0.01), 'model__alpha': 0.01, 'scaling': MaxAbsScaler()

- Mean Squared Error is: 3.395321857328653
- Root Mean Squared Error is: 1.8426399152652297
- R Squared Value is: 0.20071394356179806

This model's performance was not very accurate/reliable. The MSE/ RMSE values were quite high which suggests that there is a large amount of average error in the predictions. The R2 value shows that the features explain only a small amount of the target variable's variance. This suggests that the model did not capture the underlying patterns in the data, or relationships effectively.

Experimenting with other models: XGboost and Gradient boosting regression: - Dropping data from post publishing:

- Mean Squared Error is: 1.9302350317041748
- Root Mean Squared Error is: 1.3893289861311375
- R Squared Value is: 0.5456071585203004

Chosen model: model': GradientBoostingRegressor(learning_rate=0.01, max_depth=1), 'model__learning_rate': 0.01, 'model__max_depth': 1, 'model__min_samples_leaf': 1

This model's performance was quite good. The MSE/ RMSE values were low, showing a small average error in the predictions the model makes. The R2 value shows that the chosen features explain a meaningful portion of the target variable variance. This reveals that the model captured the underlying patterns or relationships to a fair degree.

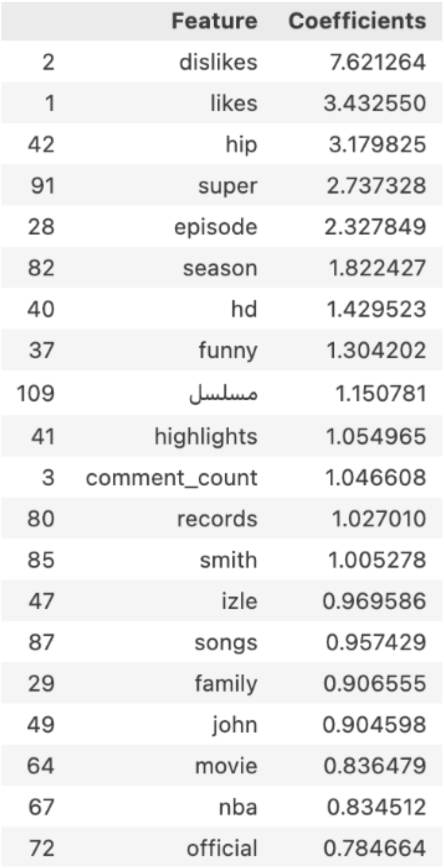
Conclusion

Final model

Baseline Lasso Regression Model with Tags only:

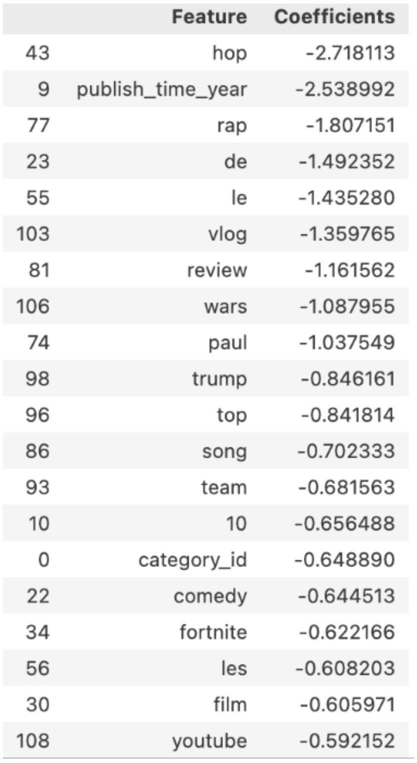
- Mean Squared Error (MSE): 1.277964599074526
- Root Mean Squared Error (RMSE): 1.1304709633929242
- R-squared (R2): 0.6991568612391986

Most predictive features:



	Feature	Coefficients
2	dislikes	7.621264
1	likes	3.432550
42	hip	3.179825
91	super	2.737328
28	episode	2.327849
82	season	1.822427
40	hd	1.429523
37	funny	1.304202
109	مسلسل	1.150781
41	highlights	1.054965
3	comment_count	1.046608
80	records	1.027010
85	smith	1.005278
47	izle	0.969586
87	songs	0.957429
29	family	0.906555
49	john	0.904598
64	movie	0.836479
67	nba	0.834512
72	official	0.784664

Least predictive features:



	Feature	Coefficients
43	hop	-2.718113
9	publish_time_year	-2.538992
77	rap	-1.807151
23	de	-1.492352
55	le	-1.435280
103	vlog	-1.359765
81	review	-1.161562
106	wars	-1.087955
74	paul	-1.037549
98	trump	-0.846161
96	top	-0.841814
86	song	-0.702333
93	team	-0.681563
10	10	-0.656488
0	category_id	-0.648890
22	comedy	-0.644513
34	fortnite	-0.622166
56	les	-0.608203
30	film	-0.605971
108	youtube	-0.592152