Analyzing YouTube Trends

Capstone Project by Liyan Cen

The problem & who might care?

As YouTuber becomes one of the most popular freelancer jobs nowadays, it's getting more challenging and competitive to achieve fame on the platform. In other words, to make money by uploading youtube videos is just becoming more difficult.

Factors might affect view counts

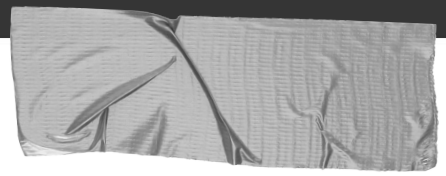
- There are many factors such as category, comment counts, likes, dislikes, publish time, trending month etc.. which can possibly affect the view counts.
- We use datasets from sources containing these factors to build supervised machine learning models to determine which factor has the most significant influence to the view counts.

Data Information

The dataset we acquire includes several months (and counting) of data on daily trending YouTube videos for US with up to 200 listed trending videos per day.

There are 37549 records on daily trending videos in US for several months.We will go through most of the fields (or columns) in the dataset to explore their relationship with view counts. The target column for this project is called “views”.

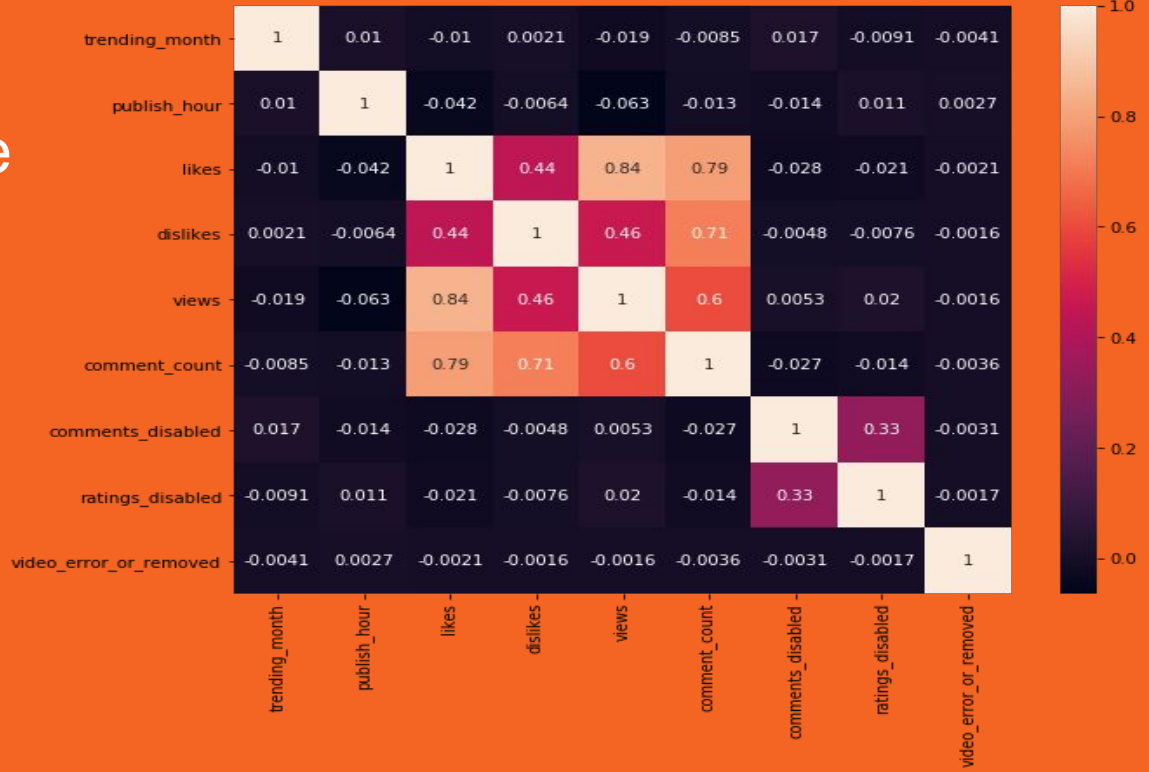
Data Exploration



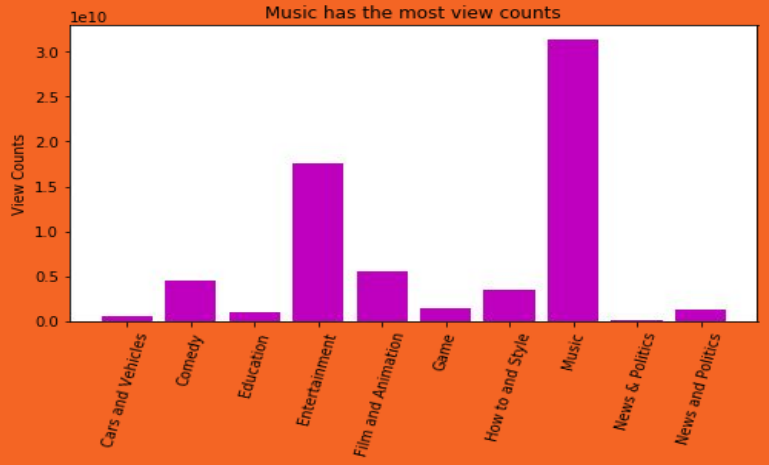
- Video Category
- Likes
- Dislikes
- View counts
- Comment counts
- Views
- Trending date
- Publish Time

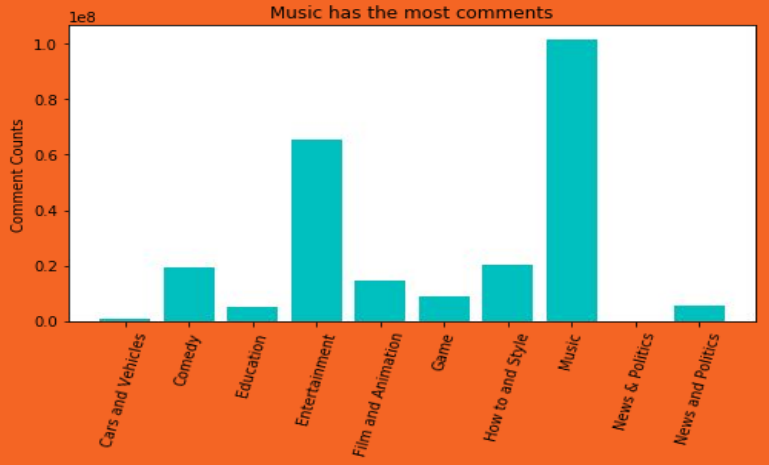
Correlation between variables

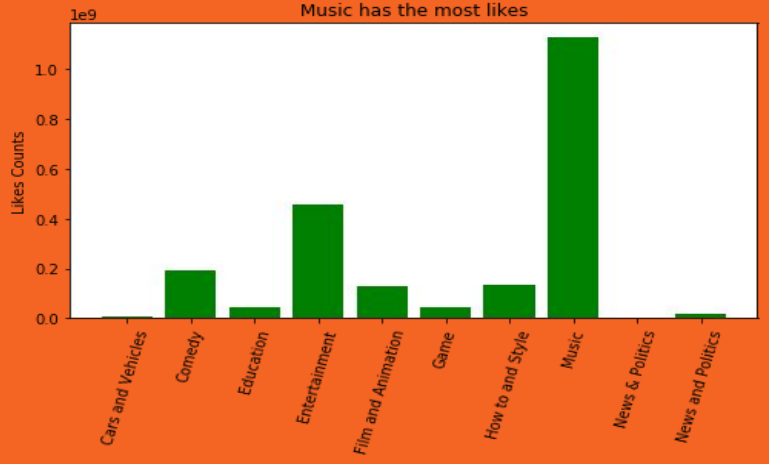
Based on the correlation matrix figure shown below, “likes”, “dislikes”, and “comment_counts” are more likely to have a correlation with the ‘views”.

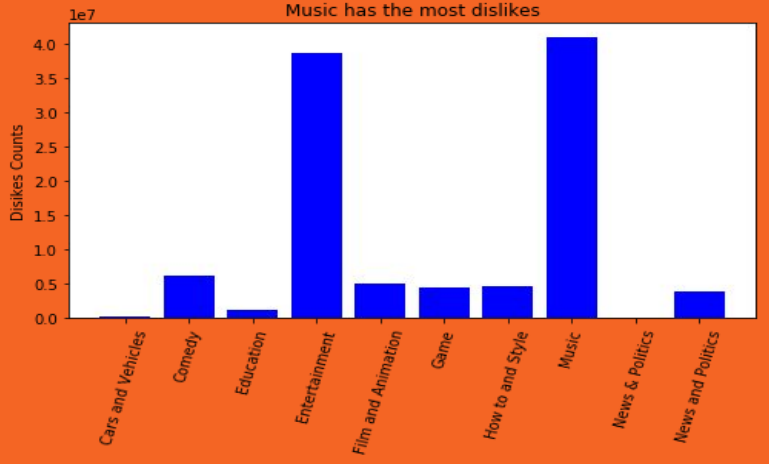


Most popular video category on youtube is “Music”





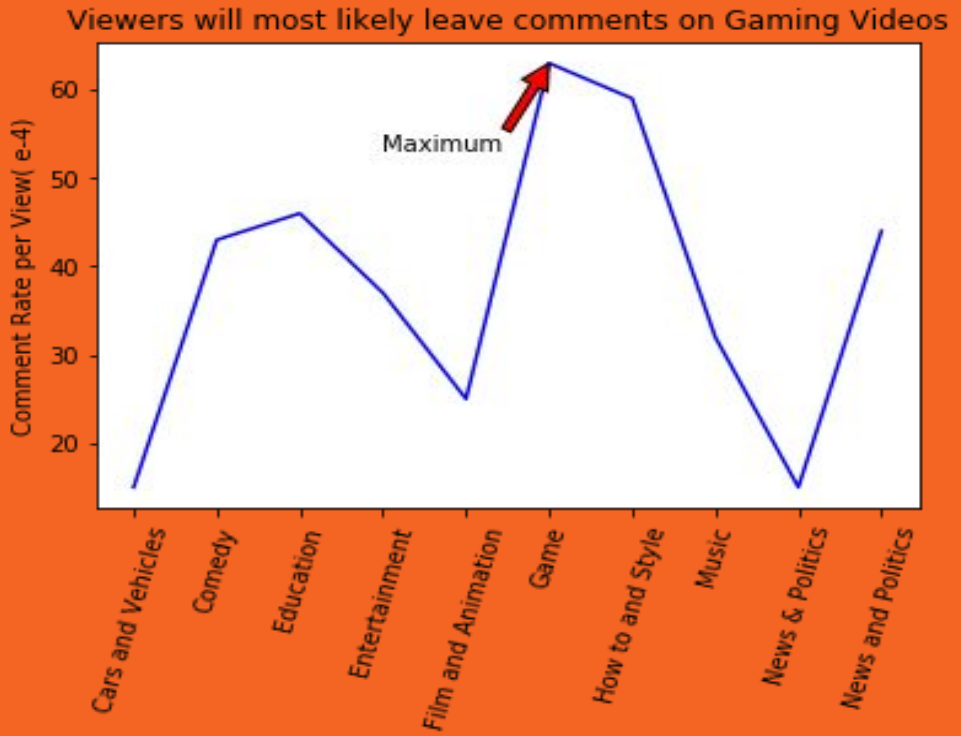




However, the results turn out differently if we compare each video category by total feature counts dividing the total view counts for each category.

Viewers will most likely leave comments on Gaming Videos

By comparing comment counts for each video category and total view counts for each category, we get the conclusion that viewers mostly like leave comments on gaming videos.

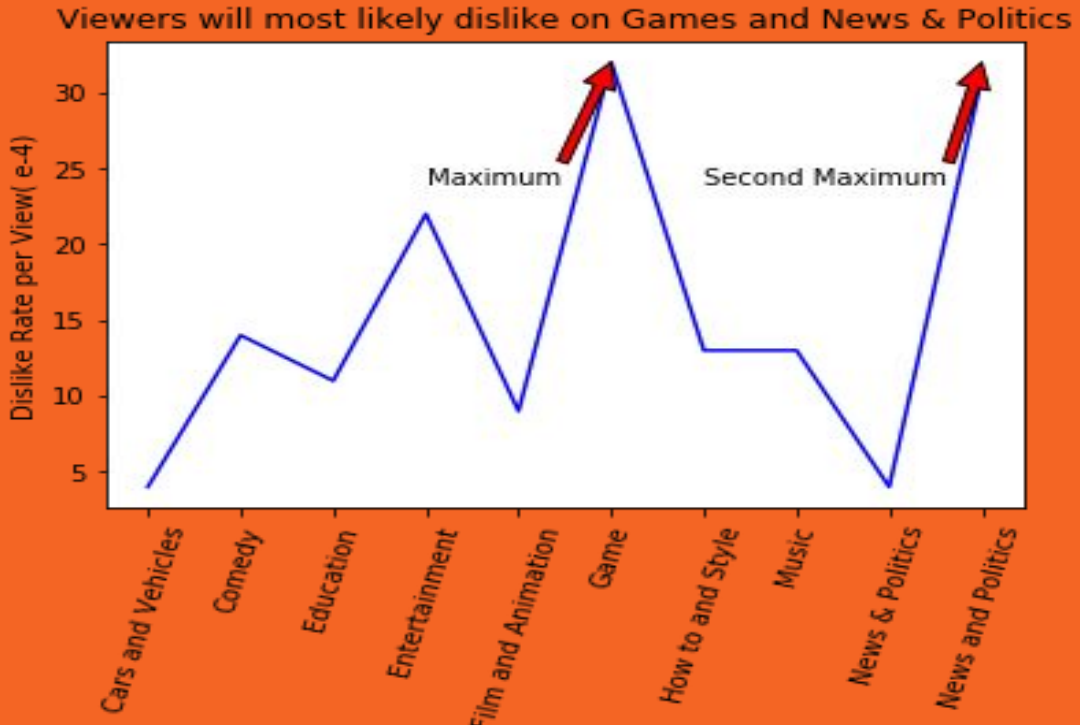


Viewers will most likely click on “like” on Comedy

By comparing like clicks for each video category and total view counts for each category, we get the conclusions that viewers mostly likely click on “like” on comedy.

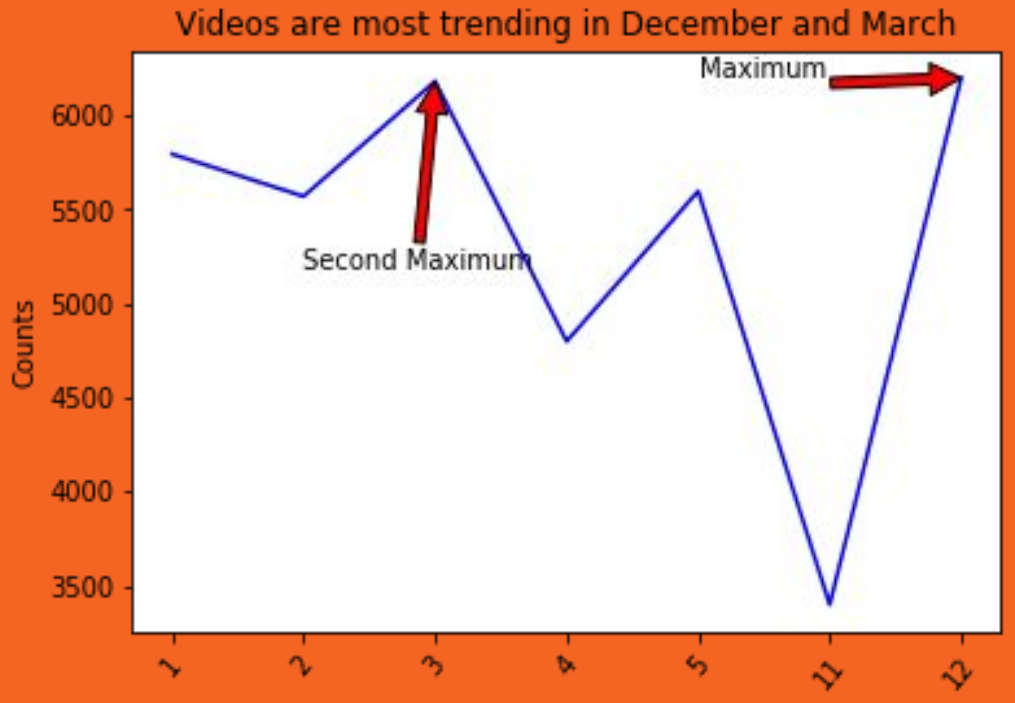
Viewers will most likely dislike on Games and News & Politics

By comparing dislike click counts for each video category and total view counts for each category, we get the conclusions that viewers mostly likely dislike on gaming videos, News and Politics.



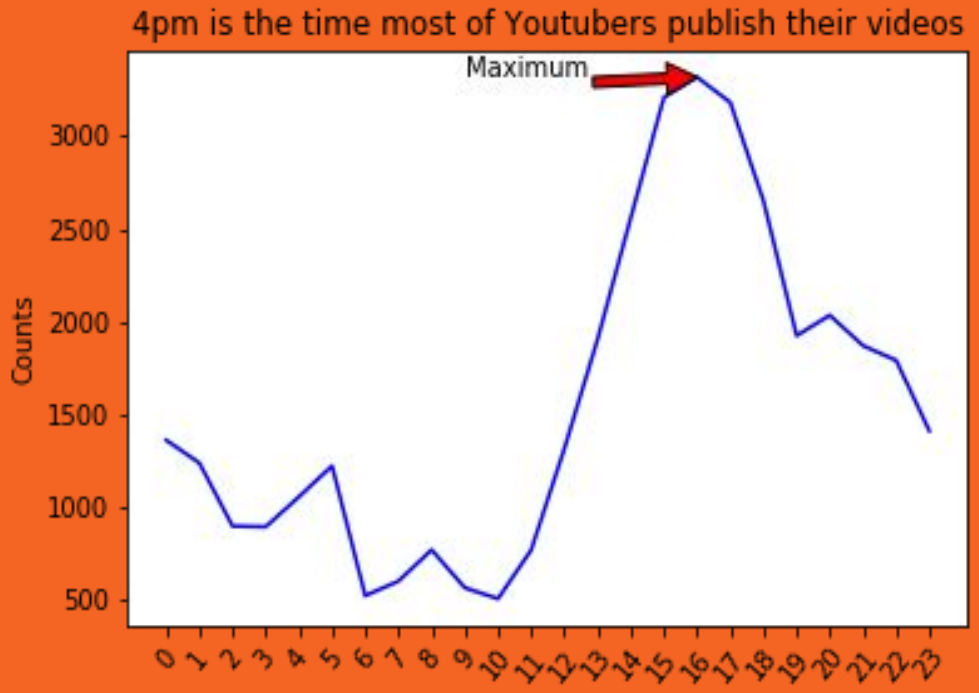
Videos are most trending in December and March

By comparing the total video counts for each month by the total trending videos overall, we get the conclusions that videos are most like to be trending in December and March.



4pm is the time most of Youtubers publish their videos

By comparing the total video counts for each hour in day by the total trending videos overall, we get the conclusions that Youtubers most likely publish their videos at 4 pm.



Modeling

We use supervised learning algorithms to build predictive models, and find out the best model based on their performance. As we mentioned earlier, we will have “comment counts”, “likes”, “dislikes”, “publish time”, “trending month” as our feature variables, and the “ views” as the predictor. We perform Linear Regression, LinearSVR, Gradient Boosting, DecisionTreeRegressor, and RandomForestRegressor, and compare their “R^2” values and “root mean square error” to figure out the best model, and best parameter.

Linear Regression

By performing cross validation training, and linear regression model, we get R^2 of 0.75, which indicates that the model explains 75% the variability of the response data around its mean. Hence, this performance is not bad.

R^2: 0.7475754728479306

Negative Root Mean Squared Error: -11158621864039.723

Support Vector Regression

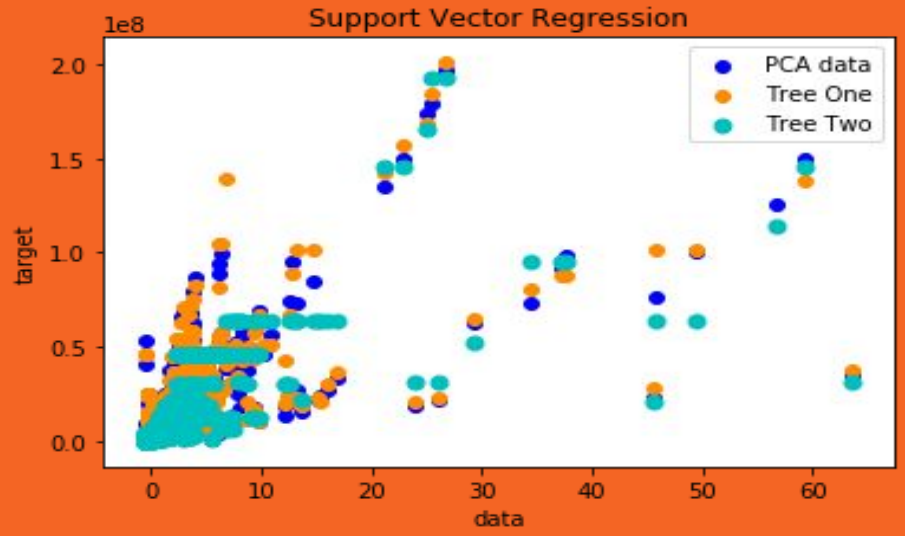
By performing the SVR linear regression model, we get R^2 of 0.62, which indicates that the model explains 62% the variability of the response data around its mean.

This is the model that gives us the lowest R^2 of all the models we perform, it might be due to the limitation of the parameter tuning.

R^2: 0.6223845633914988
Root Mean Squared Error: 4277014.810241759

Decision Tree

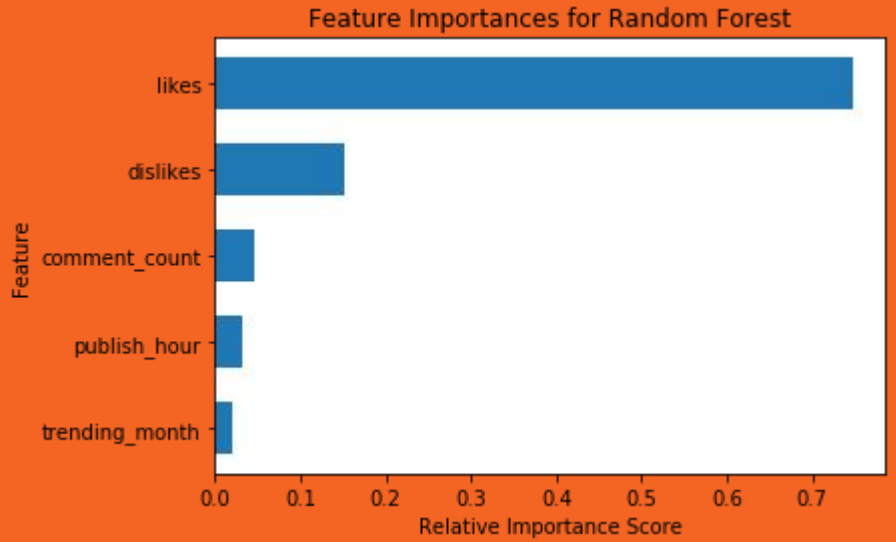
We set max_depth to be “None” and “5” for “Tree one” and “Tree two” respectively. Max_depth equalling to “None” means nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Hence, this parameter will definitely be larger than “5”. As a result, the R^2 value and the Root mean squared error are better when parameter equals to “none” than “5” .



Tree one R^2: 0.9253810156876101 # max_depth = None (no limit)
Tree two R^2: 0.843513452809086 # max_depth = 5
Tree one Root Mean Squared Error: 1901255.7957531882
Tree two Root Mean Squared Error: 2753305.4116950906

Random Forests

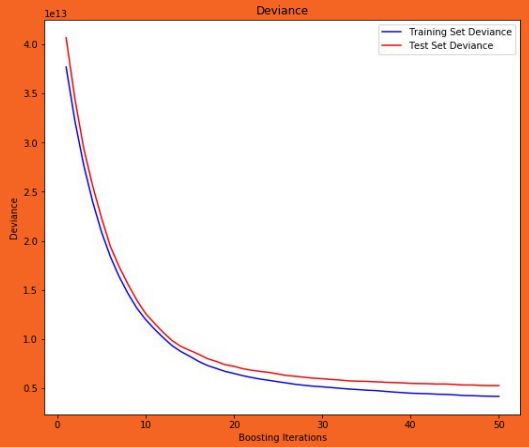
Based on the feature names ranked on the right, we get a sense that the variable “like” has the most effect in our models, which means “likes” is the feature variable that has the most significant influence to “views” in this dataset. Furthermore, by performing the random forests model, we get a best R^2 value and best root mean squared error among all model results we get.

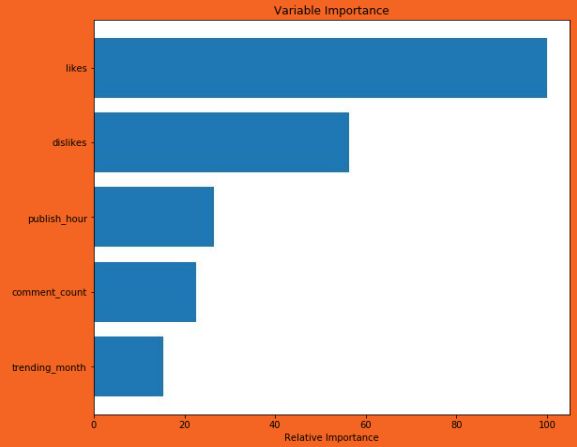


R^2: 0.9563351726789749
Root Mean Squared Error: 1454392.9404071828

Gradient Boosting

We get a R^2 value of 93%, which indicates that the model fits pretty well. And it is actually the second best model of all the models we perform. The reason why Random Forests performs a little bit stronger might be due to the fact that Gradient Boosting' training is based on last training result whereas in Random Forests, data is trained independently from the rest.





R^2: 0.9339981072293979
Root Mean Squared Error: 3197335125828.107

Hyperparameter tuning

Based on all the models listed above, it turns out that the “Random Forest” model gives us the best result.

In order to re-confirm which model is the best for our dataset, for the next two slides, due to the fact “Gradient Boosting” and “Random Forests” give us the best two performance of all, we use grid search to figure out the best hyperparameter by performing “DecisionTreeRegressor” and “RandomForestRegressor”.

DecisionTreeRegressor

param_grid = {"criterion": ["mse"], "min_samples_split": [2, 3], "min_samples_leaf": [10, 20, 30], "max_leaf_nodes": [20,40,60]}

R-Squared::0.8481297179350443

Best Hyperparameters::

{'criterion': 'mse', 'max_leaf_nodes': 60, 'min_samples_leaf': 10, 'min_samples_split': 2}

Tree one R^2: 0.8755169363559091

Tree one Root Mean Squared Error: 2455675.7321696724

After performing the cross validation for the training data, and the DecisionTreeRegressor, we get our R^2 value to be 84%. The result is not as good as the RandomForestRegressor, which may be due to the parameter limitation.

RandomForestRegressor

param_grid = {'n_estimators': [50, 100, 200], "criterion": ["mse"], "max_features": ['auto','log2', None]

R-Squared::0.9382609850453063

Best Hyperparameters::

{'criterion': 'mse', 'max_features': 'log2', 'n_estimators': 100}

Tree one R^2: 0.9656872061674027

Tree one Root Mean Squared Error: 1289270.1411325312

As a result, the RandomForestRegressor gives us the best R^2 value among all the models we have performed. Hence, it is our best hyperparameter and model for this test.

Limitations

This dataset includes several months (and counting) of data on daily trending YouTube videos for the US, so they're not the most-viewed videos overall for the calendar year.

Additionally, for this project, we use likes, dislikes, tending month, trending time, and comment counts as our feature variables to predict the view counts. However, there are many other factors which could also have significant influence on view counts like comments, share counts, subscription counts, notifications counts and etc.

Conclusions

Even though “Music” has the most “likes”, “dislikes”, “comment counts” and “views” based on the total counts of each feature in the dataset, but the rank differs if we compare each video category by total feature counts dividing the total view counts for each category. As a result, if we compare each category in ratio, the top performers on the YouTube trending list are music videos, but viewers will most likely leave comments on Gaming Videos, click “like” on Comedy, and click “dislike” on Games, News, and Politics.

On the other hand, videos are most trending in December and March. Also, most of the YouTubers like to publish their videos around 4 pm.

The feature “likes” is the factor that gives us the most significant influence on “views”. The RandomForestRegressor gives us the best R^2 value and parameter among all the models we have performed.