UNIVERSITY OF MICHIGAN

How viral will a YouTube video be?

Predicting viewership based on video metadata

YouTube

Group 25:
Chittaranjan Velambur Rajan
Jifan Li
Mengyao Wang

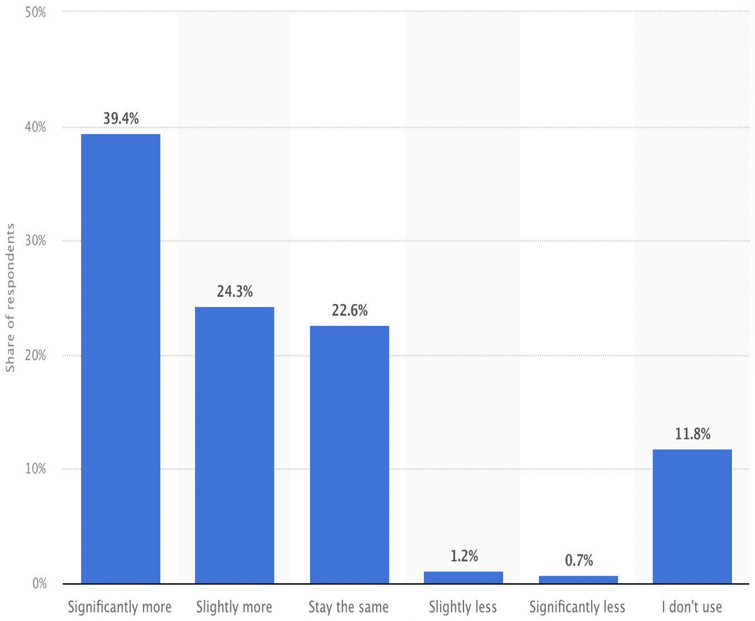
Why YouTube?

Motivation:

- Content creators want to harness this increase in viewership
- Is increased viewership just a consequence of having more free time, or something else?

Goals:

- Explore underlying trends/patterns of the data
- Identify key factors that contribute to a video's "popularity"
- Predict how many views a video will receive, given certain characteristics of the content

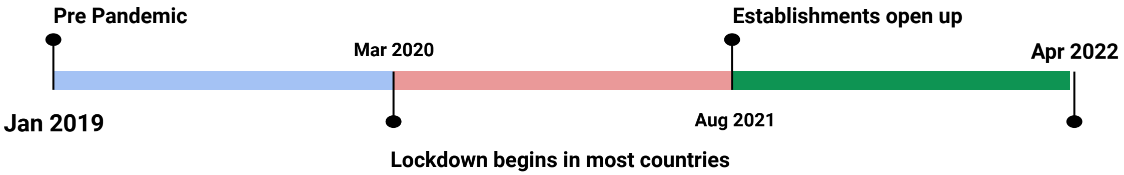


Share of social media users in the United States who believe they will use YouTube more if confined at home due to the coronavirus as of March 2020

Where is the data from?

Youtube Data API:

1. Set up software boilerplate (Access keys, API library, etc)
2. “Search” for videos in these categories:
   a. Music
   b. Comedy
   c. Entertainment
   d. News/Politics
   e. Education
   f. Science/Technology
3. Equally sample across each of these “time periods”


Pre Pandemic

Jan 2019

Mar 2020

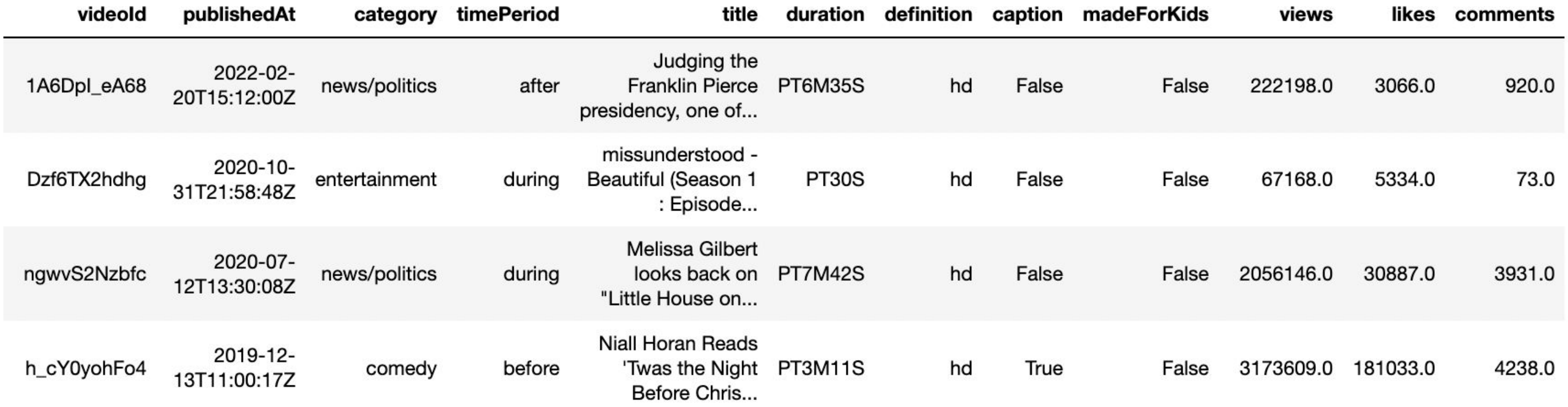
Lockdown begins in most countries

Establishments open up

Aug 2021

Apr 2022

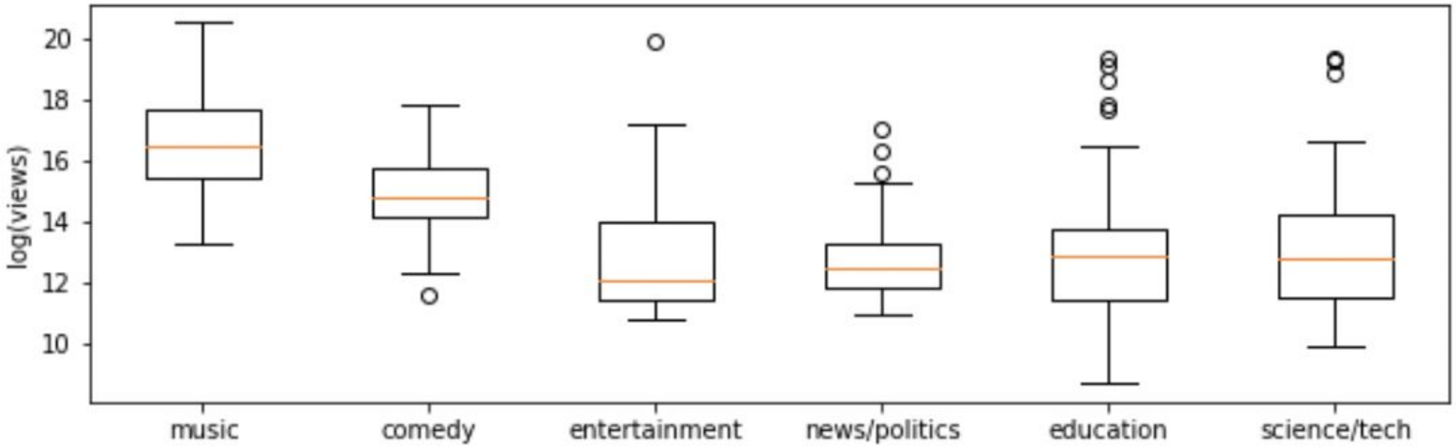
Summary of the data



videoId	publishedAt	category	timePeriod	title	duration	definition	caption	madeForKids	views	likes	comments
1A6Dpl_eA68	2022-02-20T15:12:00Z	news/politics	after	Judging the Franklin Pierce presidency, one of...	PT6M35S	hd	False	False	222198.0	3066.0	920.0
Dzf6TX2hdhg	2020-10-31T21:58:48Z	entertainment	during	missunderstood - Beautiful (Season 1 : Episode...	PT30S	hd	False	False	67168.0	5334.0	73.0
ngwvS2Nzbfc	2020-07-12T13:30:08Z	news/politics	during	Melissa Gilbert looks back on "Little House on...	PT7M42S	hd	False	False	2056146.0	30887.0	3931.0
h_cY0yohFo4	2019-12-13T11:00:17Z	comedy	before	Niall Horan Reads 'Twas the Night Before Chris...	PT3M11S	hd	True	False	3173609.0	181033.0	4238.0

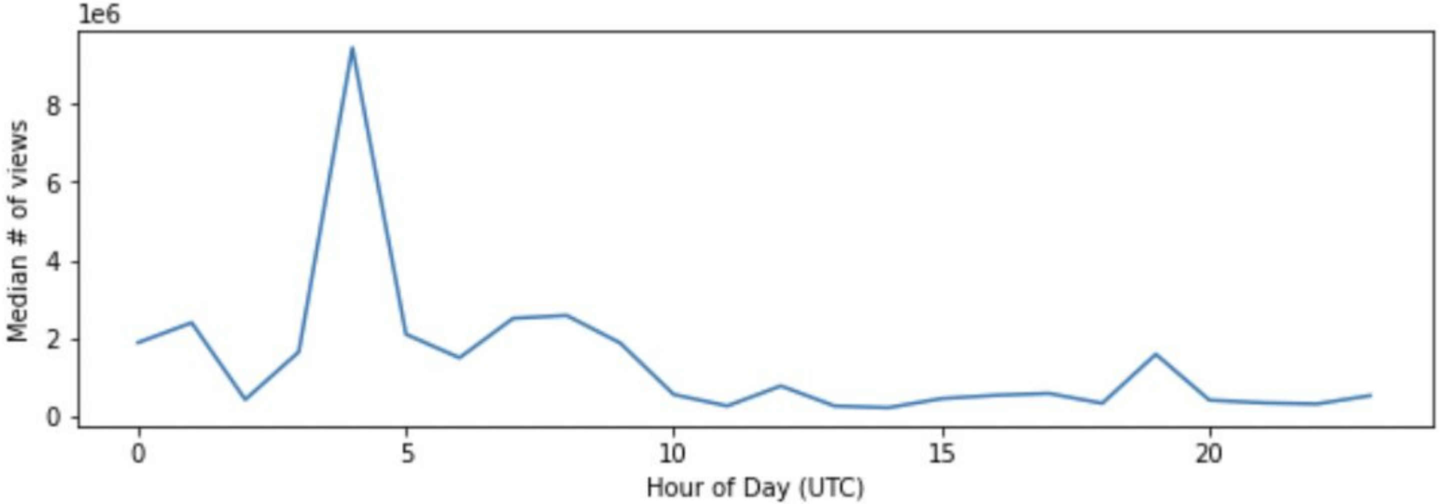
Note: Final dataset had 770 videos and 13 predictors

Exploratory Data Analysis - I



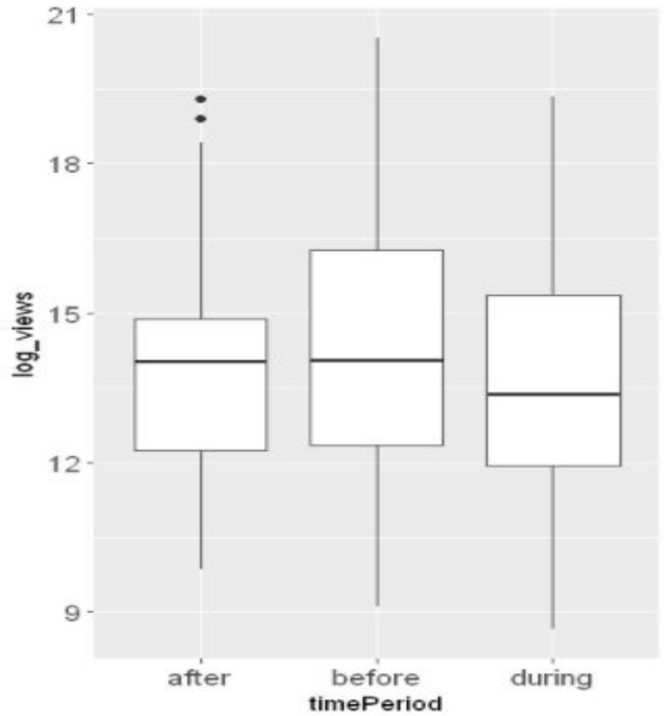
Outliers: Indirectly related to music/comedy. Eg: Katy Perry performs at Joe Biden’s inauguration is a “news/politics” video

Exploratory Data Analysis - II



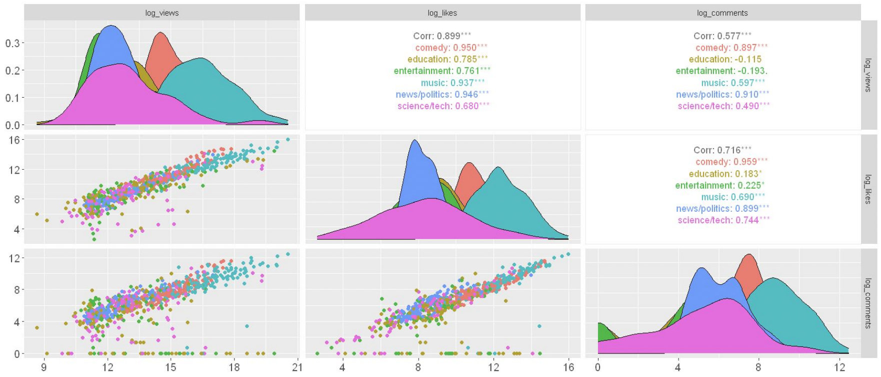
04:00 UTC corresponds to late evening and early afternoon in USA and India, the two largest source of Youtube traffic

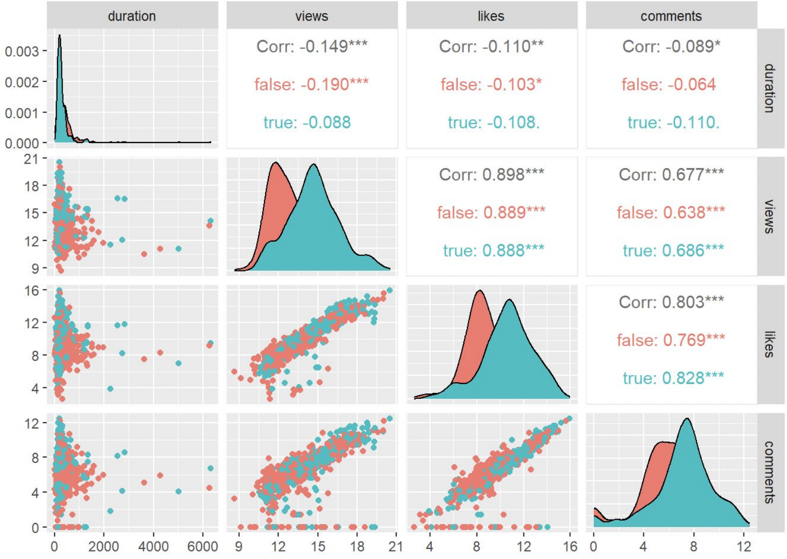
Exploratory Data Analysis - III

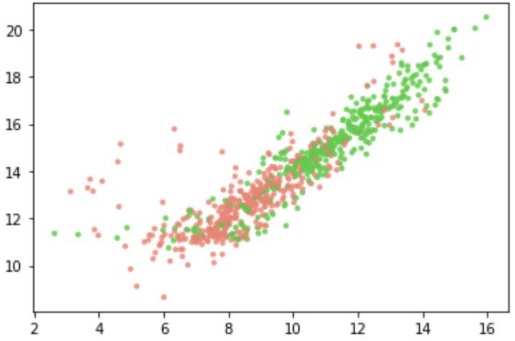


- No clear evidence of “pandemic effect” on viewership
- Inconclusive, and likely a case of insufficient data

Exploratory Data Analysis - IV





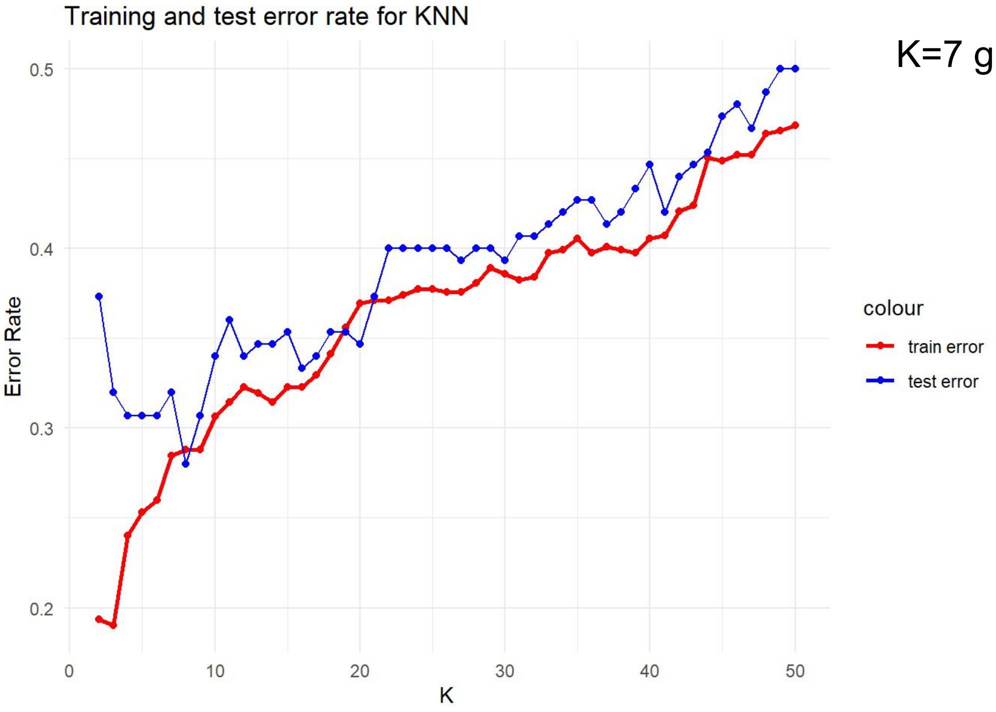


Most Promising Models

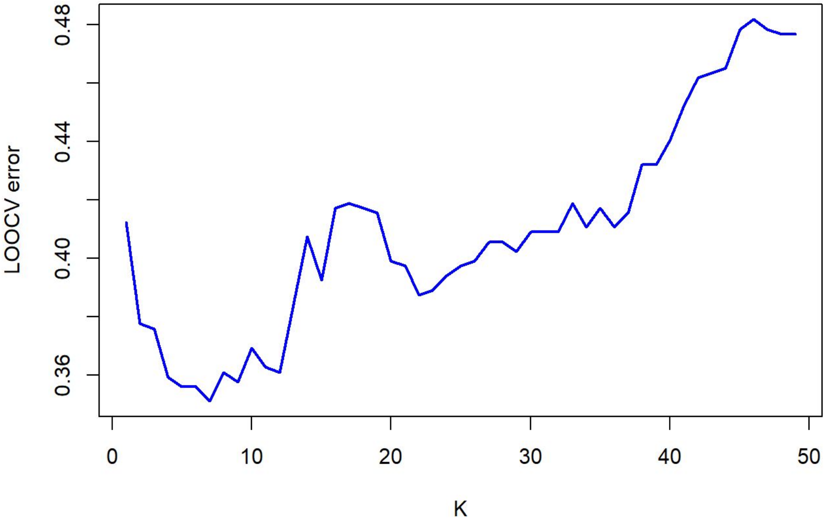
- K-Nearest Neighbors: “Similar” videos are expected to perform similarly
- Random Forest / Classification Tree: For interpretability and variable importance
- Neural Networks: Compromise interpretation for accuracy

Note: The problem was turned into a 3-way classification of “low”/“medium”/“high” number of views

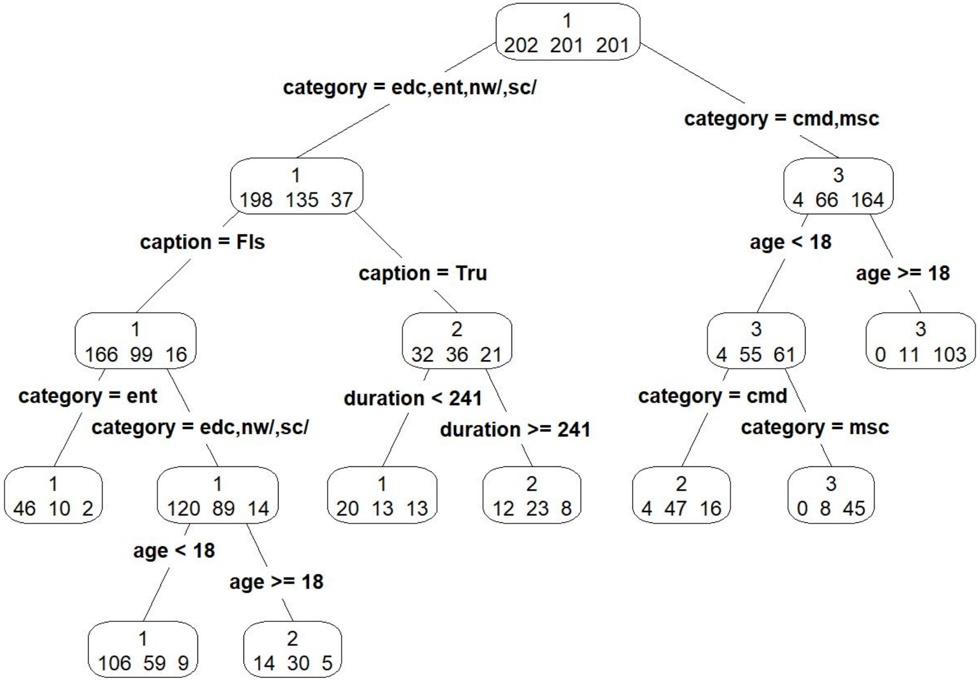
K-Nearest Neighbors



K=7 gives best accuracy of ~70% on LOOCV

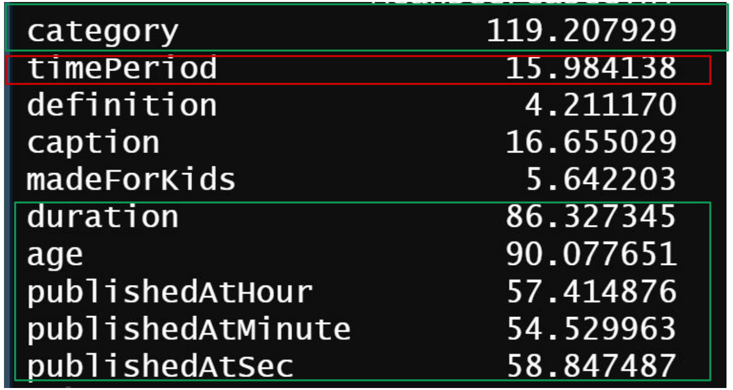


Random Forest / Classification Tree



- Slightly worse than KNN at ~68% accuracy
- Best Model
  - mtry: 3
  - nodesize: 1
  - ntree: 3000

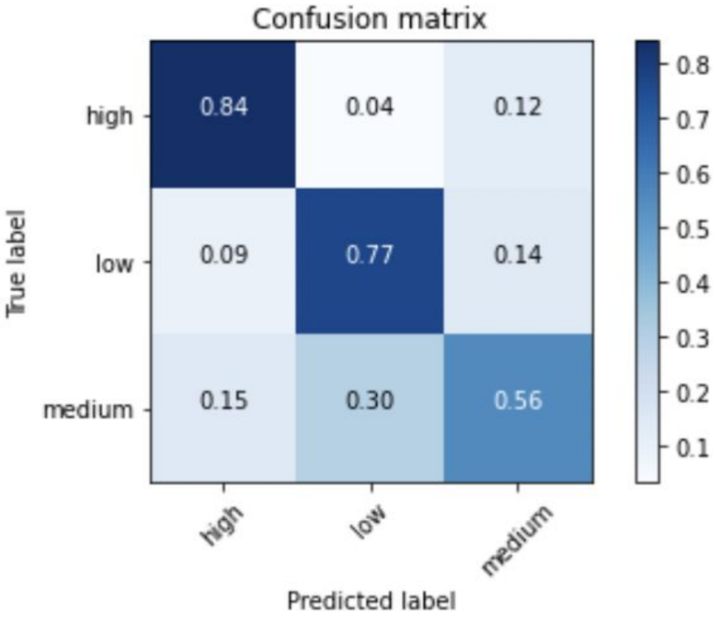
MeanDecreaseGini



```
category                  119.207929
timePeriod                 15.984138
definition                  4.211170
caption                    16.655029
madeForKids                 5.642203
duration                   86.327345
age                        90.077651
publishedAtHour            57.414876
publishedAtMinute          54.529963
publishedAtSec             58.847487
```

Neural Network

- 2 Dense Layers with 256 nodes each (ReLU)
- 1 Dropout layer (rate=0.4)
- 1 Output Layer with 3 nodes (softmax activation)



	high	low	medium
high	0.84	0.04	0.12
low	0.09	0.77	0.14
medium	0.15	0.30	0.56

Best performing model with ~72% test accuracy

Results

Model		Training Error				Test Error			
		Overall	Low	Medium	High	Overall	Low	Medium	High
LDA		0.341	0.208	0.552	0.264	0.333	0.22	0.54	0.24
Multinomial		0.343	0.183	0.582	0.264	0.333	0.18	0.56	0.26
KNN		**0.286**	**0.257**	**0.413**	**0.189**	**0.307**	**0.32**	**0.42**	**0.18**
Random Forest		**0.369**	**0.322**	**0.547**	**0.239**	**0.327**	**0.24**	**0.54**	**0.20**
GAM	Natural Splines	0.353	0.361	0.493	0.204	0.36	0.4	0.5	0.18
	Smoothing Splines	0.356	0.356	0.508	0.204	0.36	0.36	0.54	0.18
	Multiple Splines	0.359	0.396	0.478	0.204	0.36	0.36	0.54	0.18
Neural Network		**0.24**	**0.20**	**0.37**	**0.12**	**0.28**	**0.23**	**0.44**	**0.16**

Conclusions and Improvements

- We have the most error coming from predicting “medium” level videos
  - Could potentially be because of having to navigate two boundaries (low/high)
- Could potentially try regression (Neural Network had promising MSE)
  - Enables generation of a trend line of cumulative views over time
  - Predict views for the same video but vary the “age” predictor incrementally

Q & A