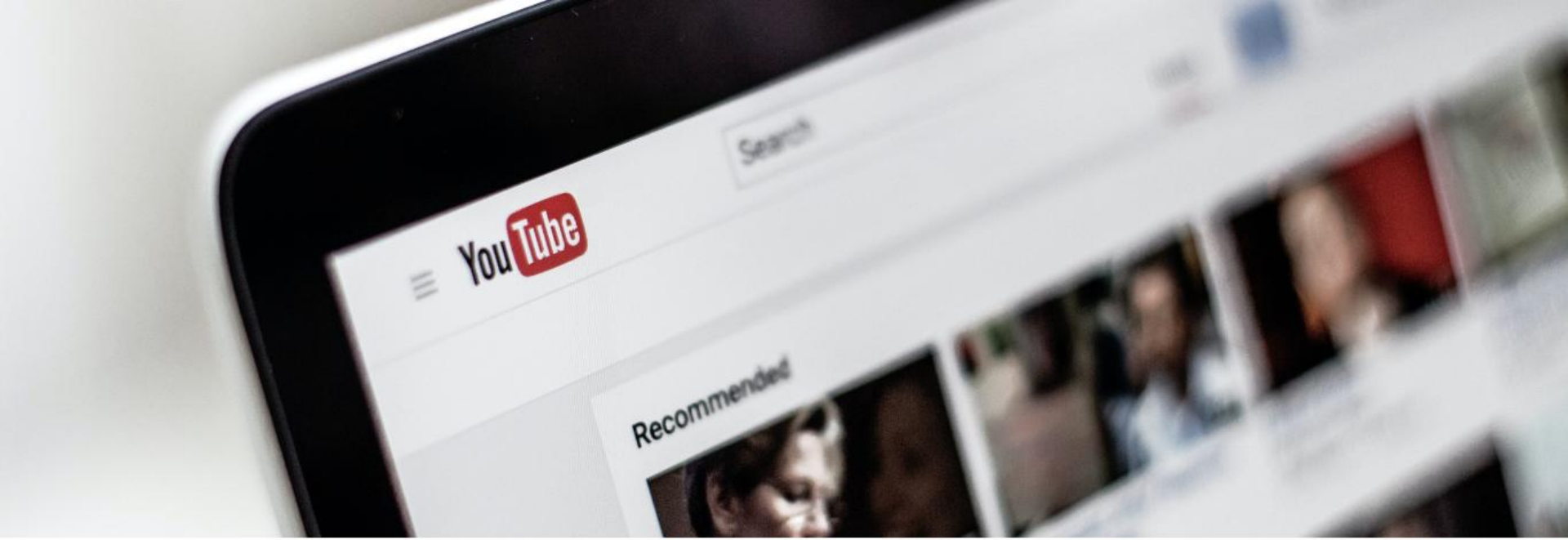


YouTube

ANALYSING YOUTUBE DATA

Team 9

Arsalan Riaz & Lakshay Bansal

MOTIVATION

We watch over 1 billion hours of YT videos a day, more than Netflix and Facebook combined. A youtuber can earn money through advertisements and given such a wide viewer base there is a lot of scope in earning money through youtube right ?

An average youtuber with 1 million subscriber makes roughly 60k USD a year. Rastko is super interested to make his own youtube channel now and earn some side money and he comes to us for his IT support.

GOALS AND OBJECTIVES

The goal is to help Rastko to have the most subscribers with the help of analytics.

-We want to understand what makes a video trending ?
-Can we predict the success of a video ?
-How can demographics help us understand what the users want to see ?




DATA

This dataset includes several months (and counting) of data on daily trending YouTube videos.

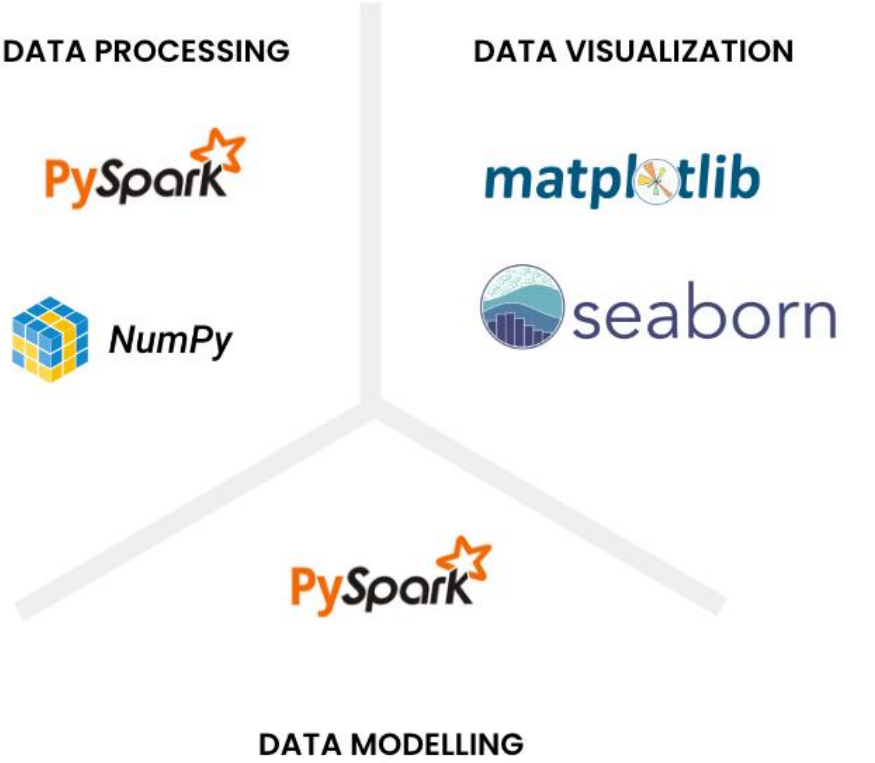
Data is included for the regions India, USA, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea, and, Japan respectively, with up to 200 listed trending videos per day.

Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

For our analysis we used data from US, India and Russia

video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	tags	view_count	likes	dislik
1FcjOMV4	Heatpacking: Last...	2021-02-22T07:30:01Z	UC3kTzVzaHQEd30rQ...	LastWeekTonight	24	21.22.02	[none]	1067147	60111	12
v84AOzj0	Best 3D Pen Art W...	2021-02-21T20:43:28Z	UCPAk4rqVIwg1NCXh...	ZHC Crafts	26	21.22.02	[none]	1047854	50662	6
KwKDdXYA	100 Days - [Minec...	2021-02-20T18:00:01Z	UC9FkeEFIGd9FXRFx...	Luke TheNotable	20	21.22.02	luke thenotable|l...	6133266	372753	7
5mt2gBjo	Amazing! Luke Bry...	2021-02-22T01:53:12Z	UCAMPco9Pqj8bI_ML...	American Idol	24	21.22.02	American Idol|sin...	790238	14267	3
cyHEb7W4	Game Theory: Did ...	2021-02-20T19:05:26Z	UCo_IB5145EVNcf8h...	The Game Theorists	20	21.22.02	fnaf|five nights ...	3248661	225780	2
Fj2D8Fe4	Deion Sanders win...	2021-02-21T23:24:44Z	UCzRdWsFjqHk1an4O...	ESPN College Foot...	17	21.22.02	deion sanders|dei...	297852	6007	
2vBHgkrg	Broner vs Santiag...	2021-02-21T06:00:12Z	UCWXYAGB9SadlL6p5...	Premier Boxing Ch...	17	21.22.02	Boxing|Combat|PBC...	820431	5233	
null	null	null	null	null	null	null	null	null	null	nu
null	null	null	null	null	null	null	null	null	null	nu
null	null	null	null	null	null	null	null	null	null	nu
null	null	null	null	null	null	null	null	null	null	nu
null	null	null	null	null	null	null	null	null	null	nu
null	null	null	null	null	null	null	null	null	null	nu
56mHdhlQ	How Do Nuclear Su...	2021-02-21T14:00:10Z	UC6107grRI4m0o2-e...	SmarterEveryDay	28	21.22.02	Smarter|Every|Day...	960432	57637	5
1SZmAghE	Rainbow Six Siege...	2021-02-21T19:59:40Z	UCBMvc6jvuTxH6TNo...	Ubisoft North Ame...	20	21.22.02	Flores|New Siege ...	294287	13574	3
ler87pwY	UNITED 328 Engine...	2021-02-21T23:04:04Z	UC88tlMjI57kf8uhP...	Captain Joe	28	21.22.02	United 328|engine...	396864	26385	3
qcf65elQ	WHO IS THE BEST C...	2021-02-21T06:11:13Z	UCPpATKqwMV-CNRNW...	Alexa Rivera	26	21.22.02	[none]	2768920	183604	25
9MBzrtns	my valentine :)	2021-02-21T21:00:09Z	UCCpeNlYK_xGS94eJ...	Clayton Bush	22	21.22.02	[none]	123607	12156	
82TcqD-A	Cards Against Hum...	2021-02-21T20:29:11Z	UCHUE4ypXKp7ZkmdW...	AmazingPh11	22	21.22.02	dan and ph11|dan ...	511276	110828	3
dZeWW91o	JAY GOT A CRUSH O...	2021-02-20T23:08:16Z	UC3xWCpnWd817HaIq...	FunnyMike	24	21.22.02	funnymike|funny m...	899525	74078	6

TOOLS USED

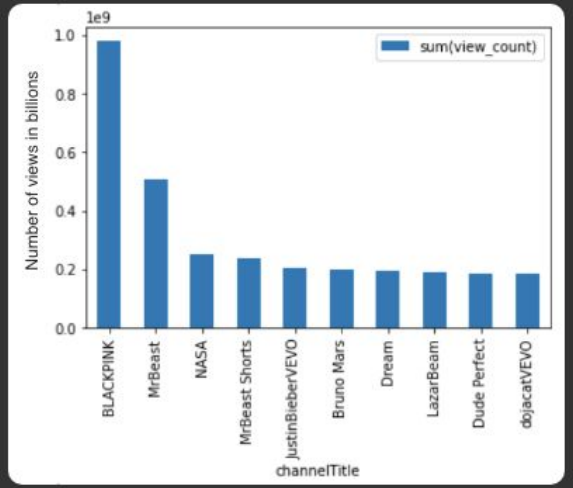

DATA PROCESSING
PySpark
NumPy
DATA VISUALIZATION
matplotlib
seaborn
PySpark
DATA MODELLING

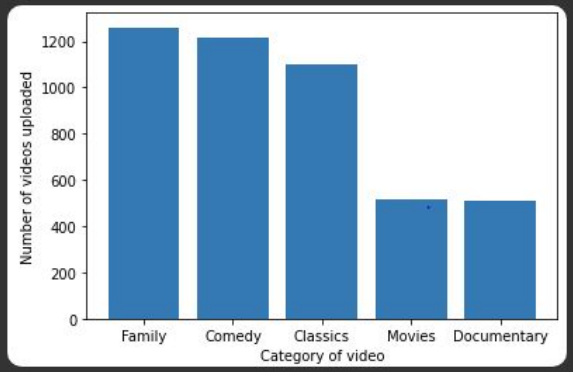
EXPLORATORY DATA ANALYSIS (USA)

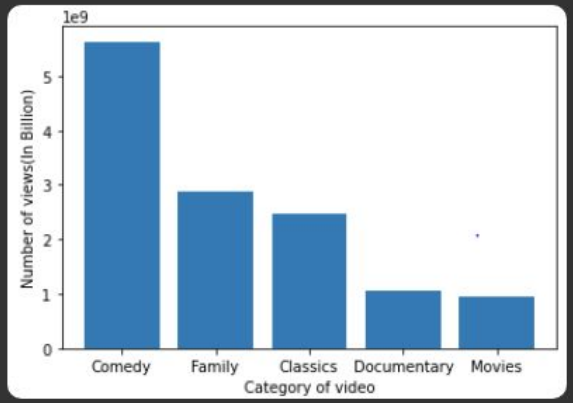
"BLACKPINK" is the most trending channel in US

Among all the categories comedy and family are the most enjoyed among users

The content that has both comedy and family are likely to trend more







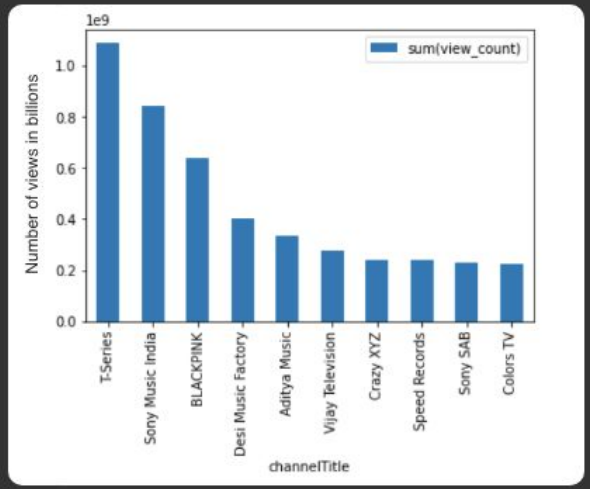
EXPLORATORY DATA ANALYSIS (INDIA)

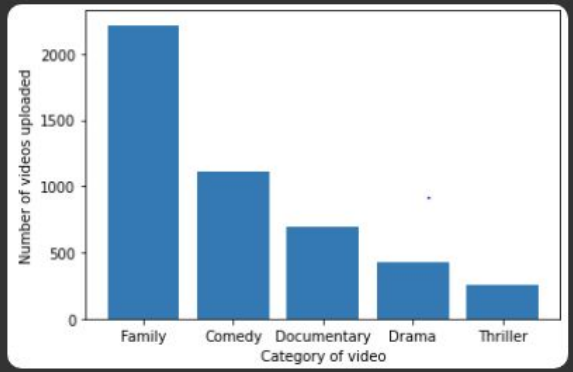
T-Series is the most trending channel in India

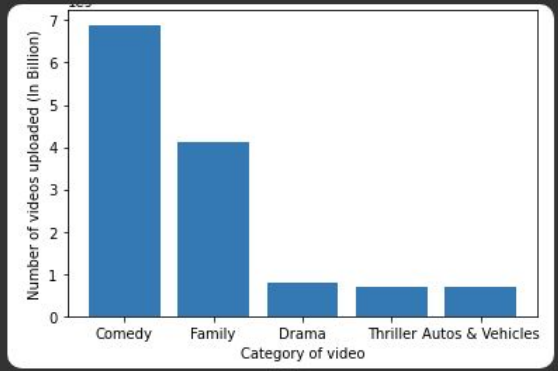
Comedy is the most trending category amongst all categories

Auto and Vehicles are amongst the top 5 trending categories

Auto and Vehicles videos should be created in more quantity as they are trending amongst top 5 categories







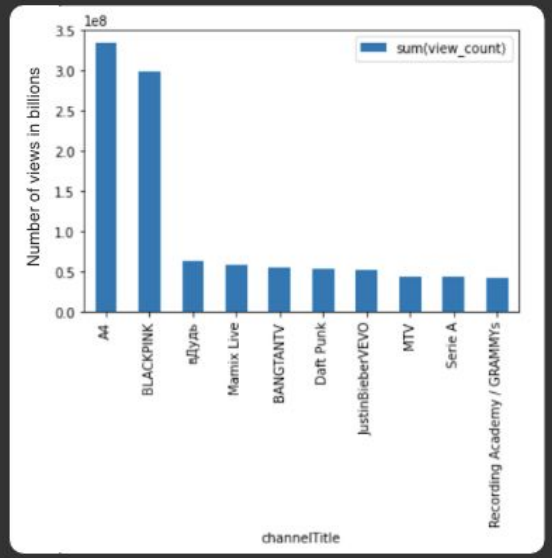
EXPLORATORY DATA ANALYSIS (RUSSIA)

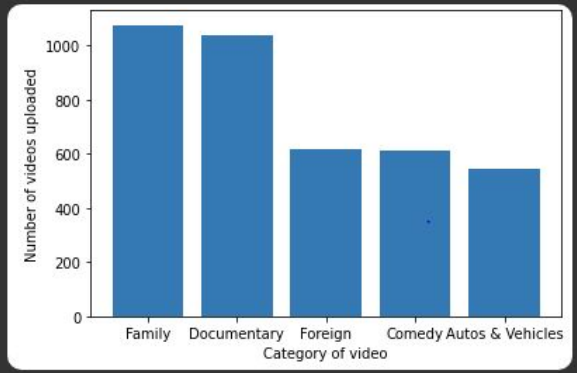
A4 is the most trending channel in Russia

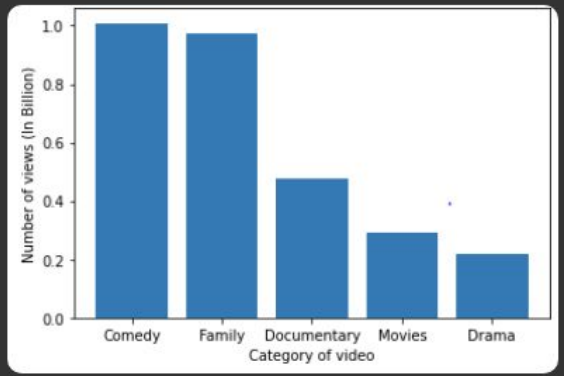
Comedy and Family are the top trending categories amongst YouTube users

Comedy and Family can be combined to produce more content

Foreign and Auto Vehicle content is not viewed as much as it is created

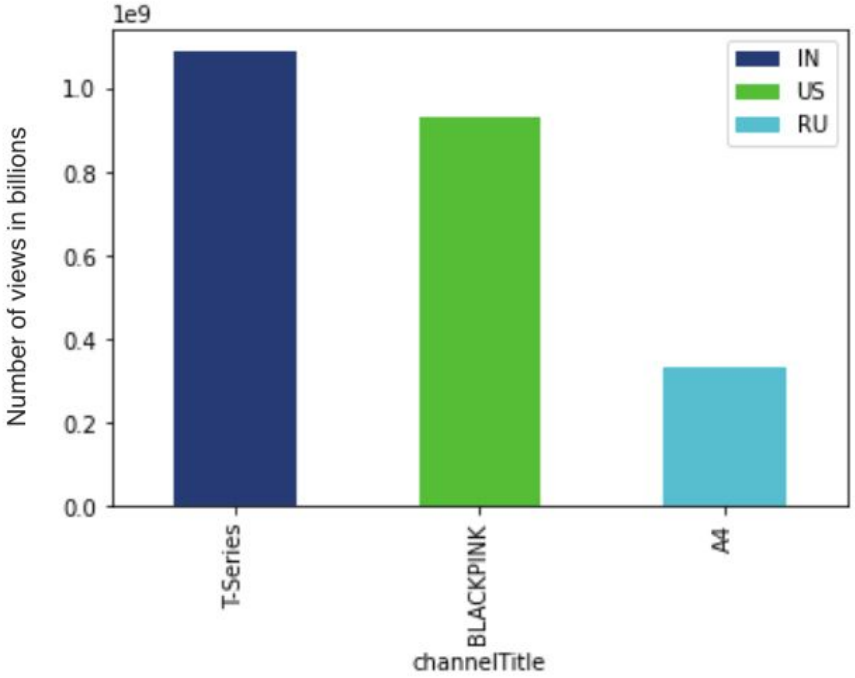






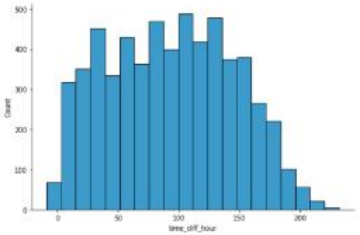
OVERALL TRENDING CHANNELS

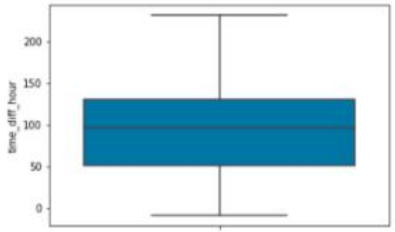
The total number of views that are received by the top viewed channels in India, Russia and United States.
Different trends are being followed in the three countries
FACT: The number of views are proportional to the population of each country



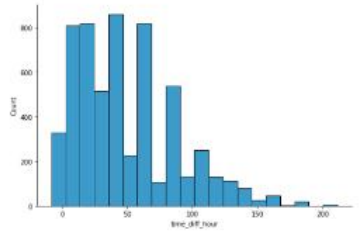
TIME TAKEN FOR VIDEOS TO GO TRENDING

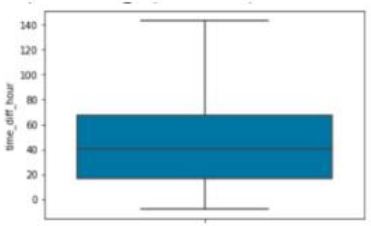
01 US - The average hours for a video to trend from the time it is published are 90



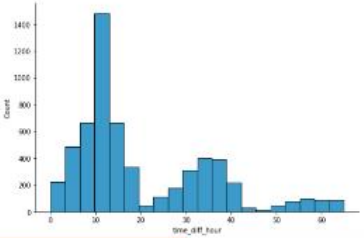


02 IN - The average hours for a video to trend from the time it is published are 40

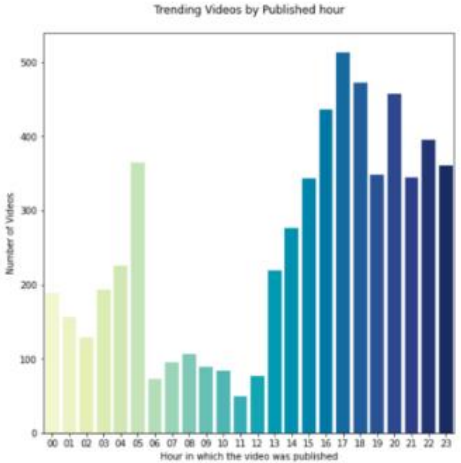




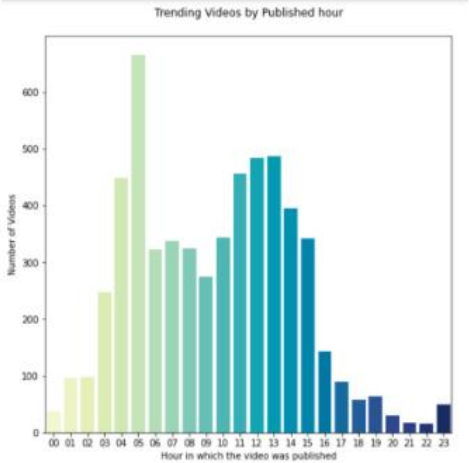
03 RU - The average hours for a video to trend from the time it is published are 15



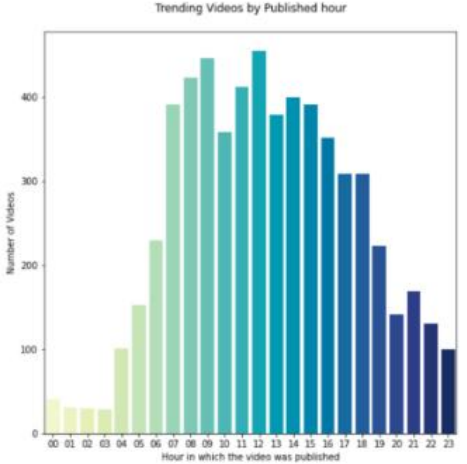
BEST TIME TO UPLOAD VIDEOS



US - The videos that received the highest views were posted on the 17th hour of the day - 5 PM CT



IN - The videos that received the highest views were posted on the 5th hour of the day - 3:30 PM IST



RU - The videos that received the highest views were posted on the 12th and 9th hour of the day - 6 PM MST

MODELLING APPROACH

Gradient boosted Tree Regression

Features used: Likes,Dislikes,Comments, CategoryId

Predicting: View count

```
[51] from pyspark.ml.linalg import Vectors
     from pyspark.ml.feature import VectorAssembler
     from pyspark.ml.clustering import KMeans
     vec_assembler = VectorAssembler(inputCols = ['categoryId','likes','dislikes','comment_count'], outputCol='features')

     final_data = vec_assembler.transform(Regression_data)
     final_data=final_data.select('view_count','features')

     (trainingData, testData) = final_data.randomSplit([0.8, 0.2])

     # Train a GBT model.
     gbt = GBTRegressor(labelCol='view_count', featuresCol="features", maxIter=8)
     gbt_model= gbt.fit(trainingData)

     # Make predictions.
     predictions = gbt_model.transform(testData)

     # Select (prediction, true label) and compute test error
     evaluator = RegressionEvaluator(
         labelCol="view_count", predictionCol="prediction", metricName="rmse")
     rmse = evaluator.evaluate(predictions)
     print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)


     Root Mean Squared Error (RMSE) on test data = 2.98671e+06
```

General Linear method Regression

```
[53] from pyspark.ml.regression import GeneralizedLinearRegression

[54] glr = GeneralizedLinearRegression(family="gaussian", link="identity", maxIter=10, regParam=0.3,labelCol='view_count',featuresCol='features')
     model= glr.fit(trainingData)
     predictions = model.transform(testData)

     # Select (prediction, true label) and compute test error
     evaluator = RegressionEvaluator(
         labelCol="view_count", predictionCol="prediction", metricName="rmse")
     rmse = evaluator.evaluate(predictions)
     print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)

     Root Mean Squared Error (RMSE) on test data = 2.17426e+06
```

MODELLING APPROACH

Used Word2Vec

```
[58] from pyspark.ml.feature import HashingTF, IDF, Tokenizer
     tokenizer = Tokenizer(inputCol="title", outputCol="Tokens_title")
     Regression_data = tokenizer.transform(Regression_data)


     from pyspark.ml.feature import StopWordsRemover
     remover = StopWordsRemover(inputCol="Tokens_title", outputCol="filtered_tokens_title")
     Regression_data=remover.transform(Regression_data)


     word2Vec = Word2Vec(vectorSize=3, minCount=0, inputCol="filtered_tokens_title", outputCol="title_embedding")
     model = word2Vec.fit(Regression_data)

     embedded_data = model.transform(Regression_data)
```

Gradient boosted Tree Regression with Title embeddings

```
[60] vec_assembler = VectorAssembler(inputCols = ['categoryId','likes','dislikes','comment_count','title_embedding'], outputCol='features')

     final_data = vec_assembler.transform(embedded_data)
     final_data=final_data.select('view_count','features')

     (trainingData, testData) = final_data.randomSplit([0.8, 0.2])

[61] gbt = GBTRegressor(labelCol='view_count', featuresCol="features", maxIter=8)
     gbt_model= gbt.fit(trainingData)

     # Make predictions.
     predictions = gbt_model.transform(testData)

     # Select (prediction, true label) and compute test error
     evaluator = RegressionEvaluator(
         labelCol="view_count", predictionCol="prediction", metricName="rmse")
     rmse = evaluator.evaluate(predictions)
     print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)

     Root Mean Squared Error (RMSE) on test data = 1.49897e+06
```

RESULTS

Model Used	Gradient Bossted Tree Regression	General Linear Regression	Gradient Boosted Tree Regression with word embeddings
RMSE	298671 ~ 298K views	217426 ~ 217K views	149897 ~ 149K views

CONCLUSION AND FUTURE WORK

DOES THIS MEAN AN END TO RASTKO'S CAREER ?

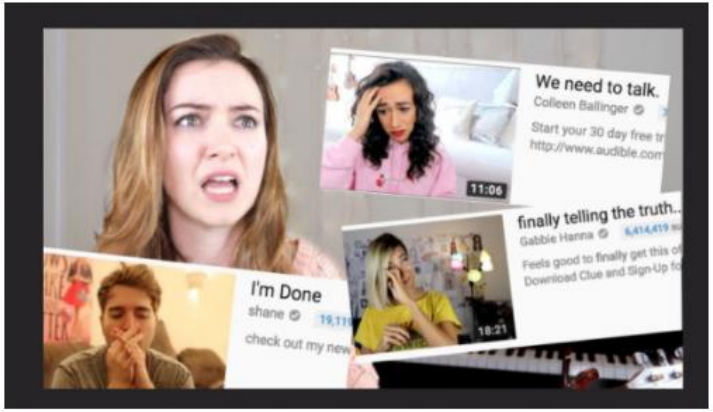
We still have stuff to try out.

1. Using external API to clickbait scores of titles

2. Using CNN on the thumbnail image to extract features from the thumbnail image can be very useful predictors.

3. Using Bert instead of word2vec for embeddings

4. Extract Topics from video titles using LDA and use that as a feature.



Example of Clickbaity titles

KEY TAKEAWAYS

1. We saw that there was a difference in the type of content users want to watch in different regions.

2. Generally uploading videos around 5pm would increase your chances of getting more views.

3. We know that number of likes and comments are helpful to determine the success of a video.

4. Adding title embedding as a feature definitely improved model performance and hence this means having eye-catchy titles help.