TEDtalks

View Count Prediction

By Hong Yee GAN, 12 November 2020

Stakeholders

"I want to scale solutions to world's most challenging problems"

"TED supports extraordinary new voices in science, arts, social justice"

"I want my ideas to reach out to more people"

"Let's spread great ideas and spark conversations"

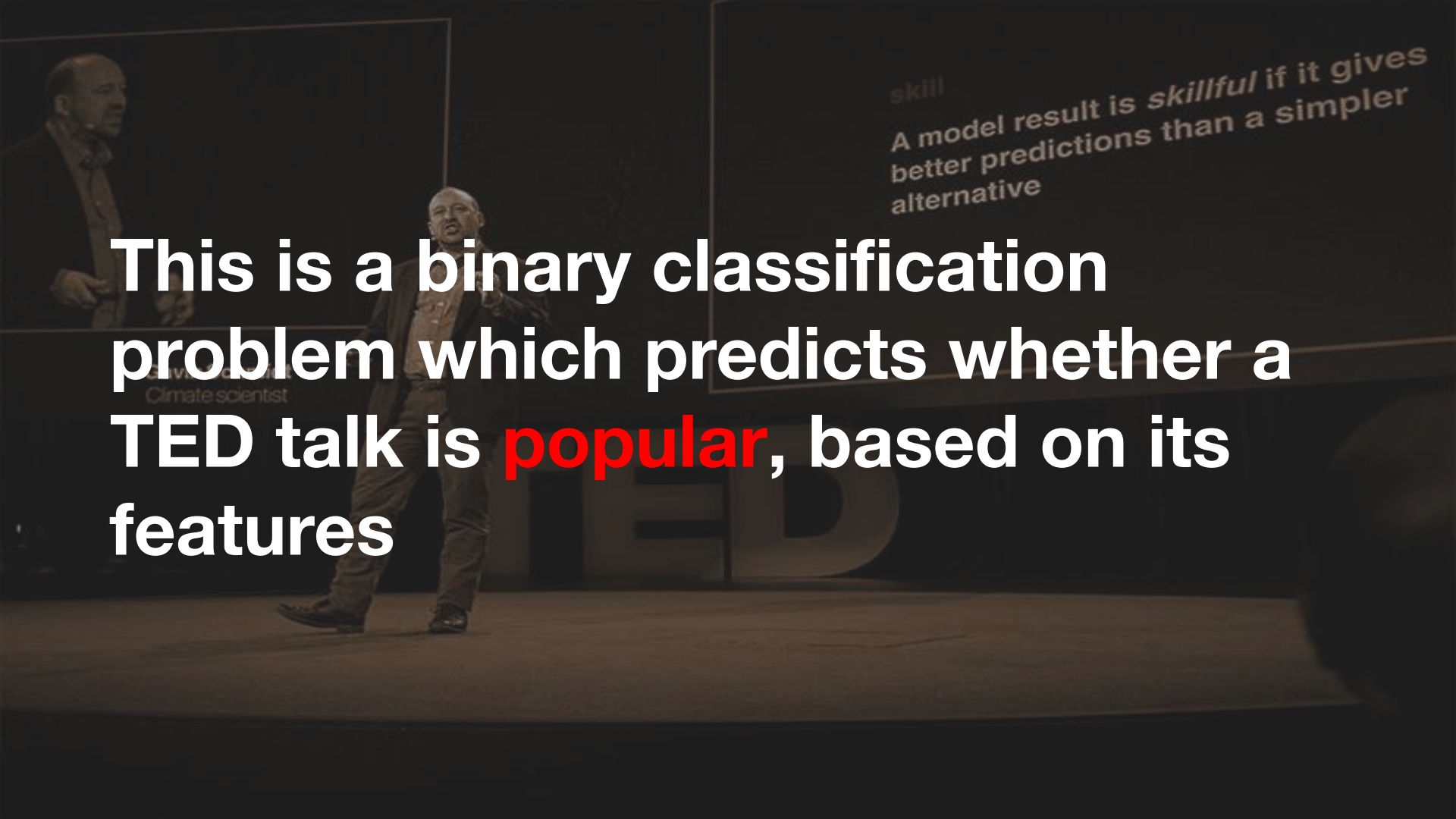
TED fundraising and marketing department

TED speakers

Me: Data Scientist working in TED

“Let’s predict which features makes a TED talk popular by developing a model”

- A data scientist working in TED would say



This is a binary classification problem which predicts whether a TED talk is popular, based on its features

TEDtalks dataset

VARIABLES

52

TRANSCRIPTS

4,609

MISSING VALUES

27,574

PUBLISHED FROM

07/2006 ~ 06/2020

Is Popular

> median

View count prediction using transcript

Model	Logistic Regression	Logistic Regression	Naive Bayes	Naive Bayes
	Tfidf	Countvectorizer	Tfidf	Countvectorizer
Train accuracy	0.788	0.977	0.755	0.751
Test accuracy	0.628	0.601	0.627	0.613
Differential	0.16	0.376	0.128	0.138

"Selected Naive Bayes Tfidf model because it has the lowest differential and high test accuracy"

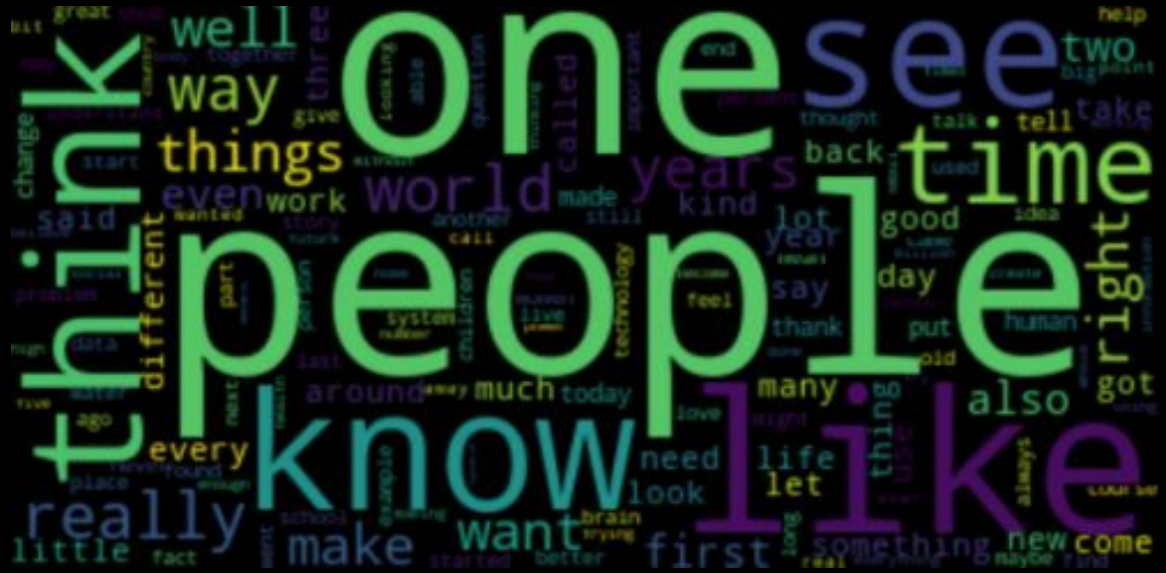
View count prediction using transcript

Dataset	Train	Test	Unseen
Accuracy score	0.755	0.627	0.631

~63% accuracy

"Selected Naive Bayes Tfidf {max features = 3000, ngram = (1,1)} model show consistency on unseen data"

Top words used in TEDtalks



"Nevertheless, I still consider this as a reasonably good model"

A model result is *skillful* if it gives better predictions than a simpler alternative

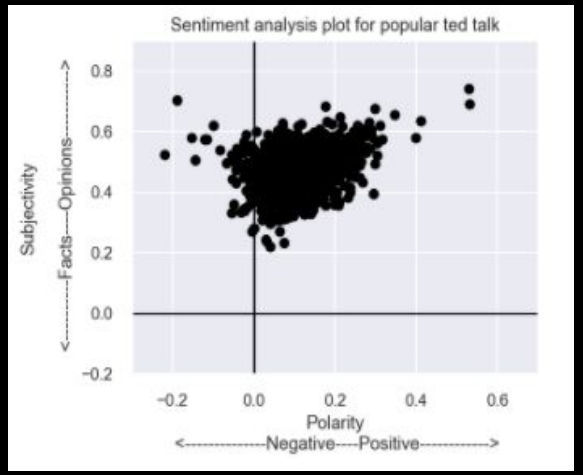
- Gavin Schmidt, Climate Scientist TED2014

Sentiment Analysis

Positive | Opinionated

Using TextBlob

Popular



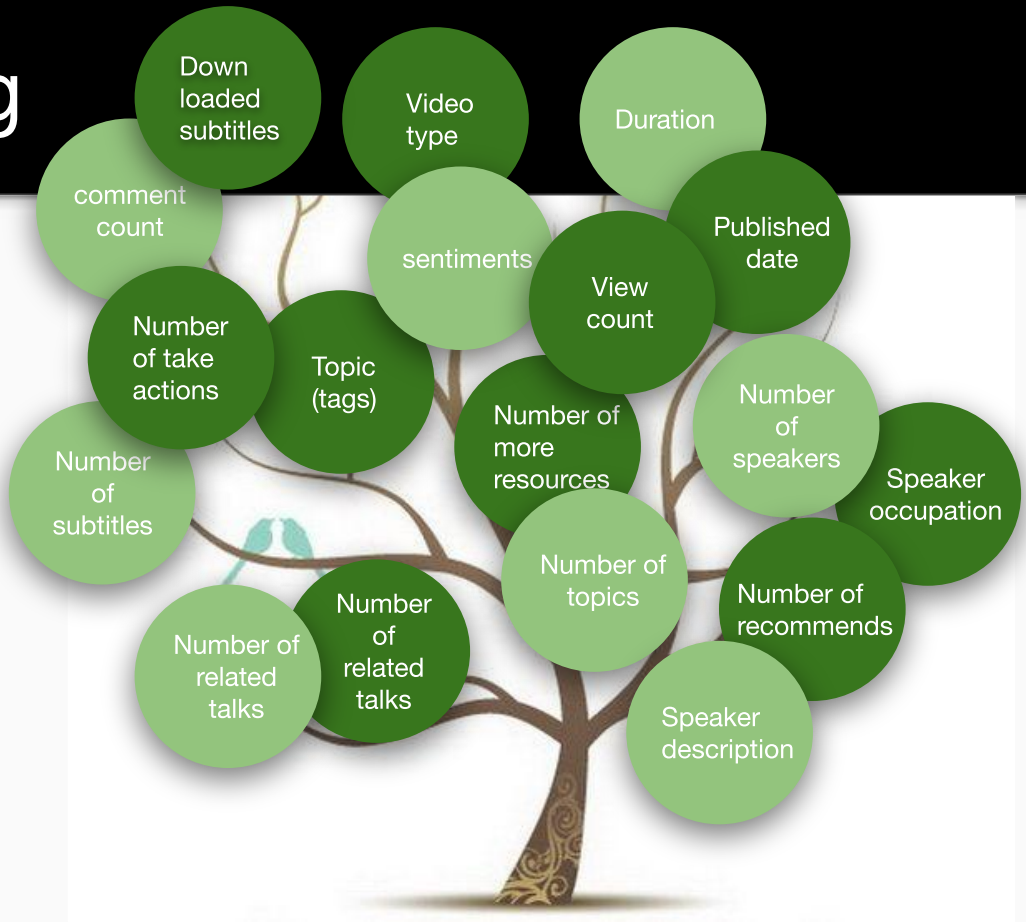
“If not the content, then what are the factors that affects view count?”

Feature Engineering

On the right are the selected features that I performed detailed EDA. Feature engineering was carried out on those more promising features.

New features:

- Video age in months
- One hot encode 30 highest view count speakers' occupations based on EDA
- One hot encode 30 highest view count topics based on EDA
- One hot encode video type



Feature Selection

reduce overfitting | increase accuracy | reduce training time

Using SelectKBest

Top 40 features using SelectKBest:

rank	feature	cumulative%	weight
1	number_of__talk__download_languages	18.2	223.349
2	comment_count	32.2	172.022
3	number_of__subtitled_videos	43.1	135.07
4	number_of__talk__more_resources	48.6	66.5778
5	work_tag	51.7	38.8593
6	number_of__talk__recommendations	54.7	36.6302
7	growth_tag	57.6	35.1469
8	personal_tag	60.4	35.1469
9	success_tag	62.9	30.3231
10	author_occupation	65.2	28.6141
11	leadership_tag	67.5	28.2233
12	psychology_tag	69.7	26.4298
13	business_tag	71.6	23.536
14	brain_tag	73.3	21.425
15	psychologist_occupation	74.9	19.4051
16	happiness_tag	76.5	19.3618
17	number_of__talks__take_actions	78	18.985
18	age_months	79.4	16.9288
19	ted salon talk partner _video	80.7	15.6402
20	duration	81.8	14.1384
21	self_tag	82.9	13.0351
22	ted institute talk_video	83.9	12.2438
23	communication_tag	84.8	11.2523
24	subjectivity	85.7	10.9645
25	ted stage talk_video	86.5	10.2511
26	ted original_video	87.3	9.5948
27	engineer_occupation	88	9.2726
28	researcher_occupation	88.7	8.3277
29	artist_occupation	89.4	8.2872
30	teaching_tag	90	7.4645
31	writer_occupation	90.6	7.4611
32	technology_tag	91.2	7.0033
33	best web_video	91.7	6.8642
34	tedx talk_video	92.3	6.6644
35	body_tag	92.8	6.5012
36	philosopher_occupation	93.3	6.2216
37	comedian_occupation	93.8	6.1864
38	original content_video	94.3	6.1769
39	comedy_tag	94.8	6.0387
40	entrepreneur_tag	95.3	5.8165

View count prediction using features

Model	Support Vector Classifier	RandomForest	KNeighbors Classifier	Logistic Regression
Train accuracy	0.756	0.771	0.739	0.749
Test accuracy	0.745	0.747	0.696	0.742
Differential	0.011	0.024	0.043	0.007

"Selected logistic regression model as it has the lowest differential and high test accuracy"

View count prediction using features

Dataset	Train	Test	Unseen
Accuracy score	0.749	0.742	0.748

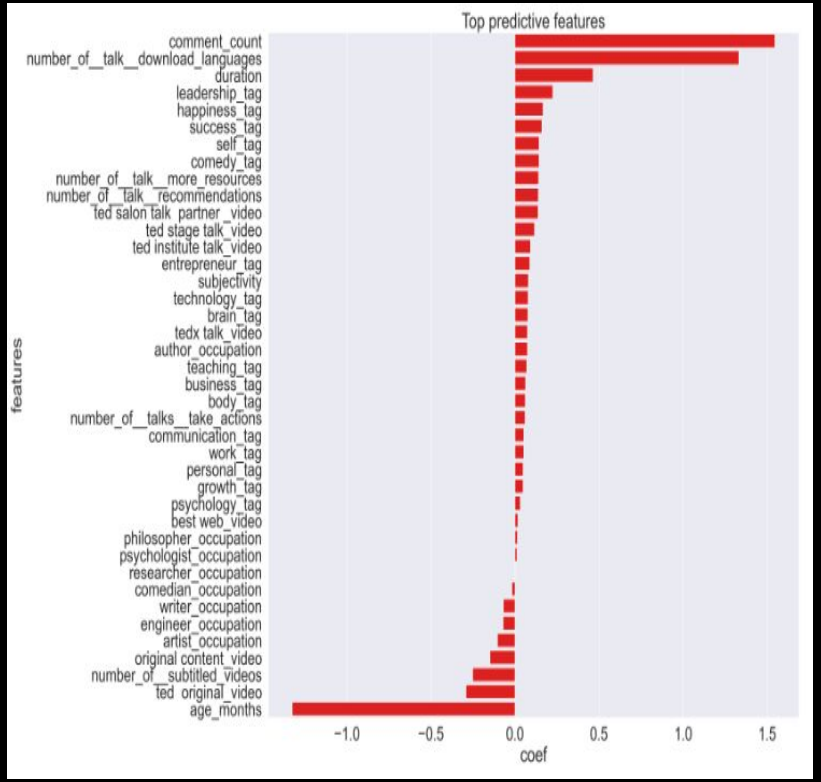
~75% accuracy

0% overfit

Logistic Regression {C = 1, penalty = l2, random state=42}

Conclusion

“ A popular TED talk is highly commented, between 8 to 16 min in duration, high number of downloaded languages, recently presented in person, based on a leadership topic ”



Recommendations

For Data Scientist in TED:

- Since the top predictive feature is comment count and it can only be generated after talk is published, we can expand the model to study what are the features that will spur comments
- Model can be expanded to include other features eg number of words in transcripts, talking speed and gender

For TED fundraising and marketing

- Publish less animated video
- Invite leaders
- Increase number of embedded subtitles for downloading

For TED speakers:

- Make adjustment to presentation style.

Recommender

Content based filtering using tfidf and cosine similarity

How it works?

1) Input the talk id, number of recommends required
2) Talk name will appear
3) Recommendations of similar talks is then generated

Improvements needed

1) Evaluate performance statistically
2) Includes other attributes
3) Explore how to deploy this recommender

Example 1

```
recommend(1042,3)
```

```
Recommending 3 ted talks similar to: The power of vulnerability
https://www.ted.com/talks/brene_brown_the_power_of_vulnerability

------------------------------------------------------------
You may also like to view: An art made of trust, vulnerability and connection (score:0.21840989708732303)
https://www.ted.com/talks/marina_abramovic_an_art_made_of_trust_vulnerability_and_connection

You may also like to view: The power of time off (score:0.18263010476165611)
https://www.ted.com/talks/stefan_sagmeister_the_power_of_time_off

You may also like to view: How to understand power (score:0.1778930987294059)
https://www.ted.com/talks/eric_liu_how_to_understand_power
```

Example 2

```
# recommendation for talk id 20319
recommend(20319,3)
```

```
Recommending 3 ted talks similar to: How do cigarettes affect the body?
https://www.ted.com/talks/krishna_sudhir_how_do_cigarettes_affect_the_body

------------------------------------------------------------
You may also like to view: How rollercoasters affect your body (score:0.421405328448158)
https://www.ted.com/talks/brian_d_avery_how_rollercoasters_affect_your_body

You may also like to view: What you should know about vaping and e-cigarettes (score:0.1748056453106719)
https://www.ted.com/talks/suchitra_krishnan_sarin_what_you_should_know_about_vaping_and_e_cigarettes

You may also like to view: Own your body's data (score:0.16312576817621094)
https://www.ted.com/talks/talithia_williams_own_your_body_s_data
```

Project Achievement

- View count prediction using transcripts ~63% accuracy
- View count prediction using features ~75% accuracy with 0% overfit
- Sentiment Analysis All talks are positive and opinionated. This aligned with TED vision
- Created a content based recommender Reasonable recommendations made. Further improvement needed

"Remember to say thank you"
~Laura Trice TED2008