INSY 695 - Group 3

What makes a popular TED Talk?

Ram Babu

Sophie Courtemanche-Martel

Arnaud Guzman-Annès

Duncan Wang

Jules Zielinski

About us!

TED Ideas worth spreading

Ram Babu
BA & Sc. ‘20
Role: Data Scientist

Arnaud Guzman-Annes
M.Eng ‘20
Role: Project Manager/SME

Sophie Courtemanche-Martel
BSc. ‘19
Role: Data Scientist / UX/UI specialist

Duncan Wang
BA & Sc. ‘20
Role: Product Manager/ Data Analyst

Jules Zielinski
BA ‘20
Role: Business Analyst

Context

TED talks are influential talks hosted by TED Conferences LLC. Since 1990, TED has hosted annual conferences where speakers in fields ranging from tech, to science, to education have been invited to come and share their ideas in the form of a short talks.

TED currently awards $1 million to one winner from each year's speaker pool, with the goal of using this prize to fund a project related to an innovative idea or project that the winner speaker has proposed. The TED style-conference model has become largely successful, and has since generated several independent spinoffs and partnership conferences which have taken place in over 130 countries.

As video recordings of TED talks have garnered over 1 billion views to date, it is evident that ***TED represents a significant platform and opportunity for anyone with a powerful mission to raise awareness and attention to their work.***

Objective

WHAT: The goal of this project is to predict the **number of views** a TED talk can expect to receive, and better understand what **attributes** are associated with popular TED Talks.

HOW: We carried out a complete study on a collection of all talks featured on TED's website from 2006 to 2017 by performing EDA, creating supervised prediction models, and unsupervised clustering models to support the analysis.

WHY: This work will allow future speakers to understand the formula for creating a unique show that will bring them views and which will therefore be profitable for both the speaker, the company it is representing, and the TED enterprise.

In other words, our team will be able to list out the different attributes that makes a given TED talk more popular than others, and leverage them to increase engagement.

Data Overview

TED Ideas worth spreading

Sample of original data

	name	title	description	tags	ratings
0	Ken Robinson: Do schools kill creativity?	Do schools kill creativity?	Sir Ken Robinson makes an entertaining and pro...	['children', 'creativity', 'culture', 'dance',...	[{'id': 7, 'name': 'Funny', 'count': 19645}, {...
1	Al Gore: Averting the climate crisis	Averting the climate crisis	With the same humor and humanity he exuded in ...	['alternative energy', 'cars', 'climate change...	[{'id': 7, 'name': 'Funny', 'count': 544}, {'i...
2	David Pogue: Simplicity sells	Simplicity sells	New York Times columnist David Pogue takes aim...	['computers', 'entertainment', 'interface desi...	[{'id': 7, 'name': 'Funny', 'count': 964}, {'i...
3	Majora Carter: Greening the ghetto	Greening the ghetto	In an emotionally charged talk, MacArthur-winn...	['MacArthur grant', 'activism', 'business', 'c...	[{'id': 3, 'name': 'Courageous', 'count': 760}...
4	Hans Rosling: The best stats you've ever seen	The best stats you've ever seen	You've never seen data presented like this. Wi...	['Africa', 'Asia', 'Google', 'demo', 'economic...	[{'id': 9, 'name': 'Ingenious', 'count': 3202}...

Basic steps

1. Setup : Software version, root directories
2. Get the data (from Kaggle, stored on GitHub)
3. Retrieving basic information
4. Formatting and extracting desired variables of targeted columns

Numerical Attributes

- Talk duration
- Main speaker
- Speaker occupation
- Number of comments
- Number of translated languages
- Views

Categorical Attributes

- Talk name
- Main speaker
- Speaker occupation
- Event category
- Topic category (tags)
- Related talks

Date Features

- Film date
- Published date
- Speaker occupation

Hypothesis

"We hypothesized that due to the wide range of data types and distributions, a ***tree-based algorithm such as Gradient Boosting Regressor would perform most optimally at prediction.***

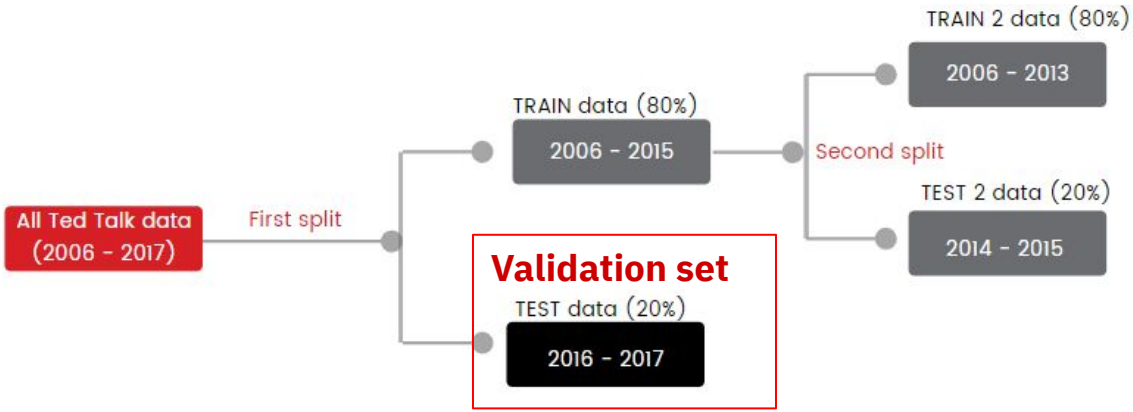
We further hypothesized that ***TED talk viewers are drawn to or away from a talk based on its content****, such as its topic, theme, or speaker, rather than other attributes such as how long it is or when it was released."*

Methods - Part 1: Split data

Time based cross validation approach: the data itself is not time-series but still has a time dimension aspect.

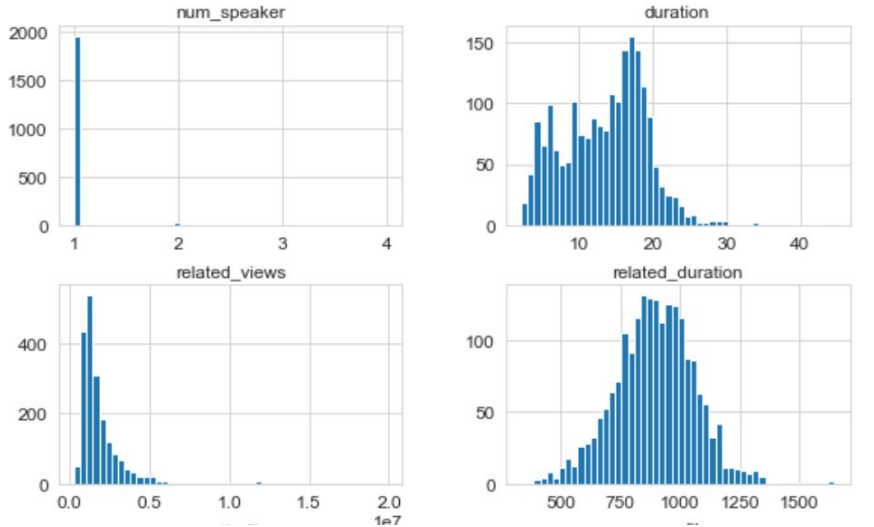
Useful approach to to build statistically robust models and follow up with time based cross validations to evaluate the performance of the final model.

The basis of the splits for the train and test data will be the year of release, with consideration with the amount of talks in each year to ensure a 80% and 20% split for training and test, respectively.

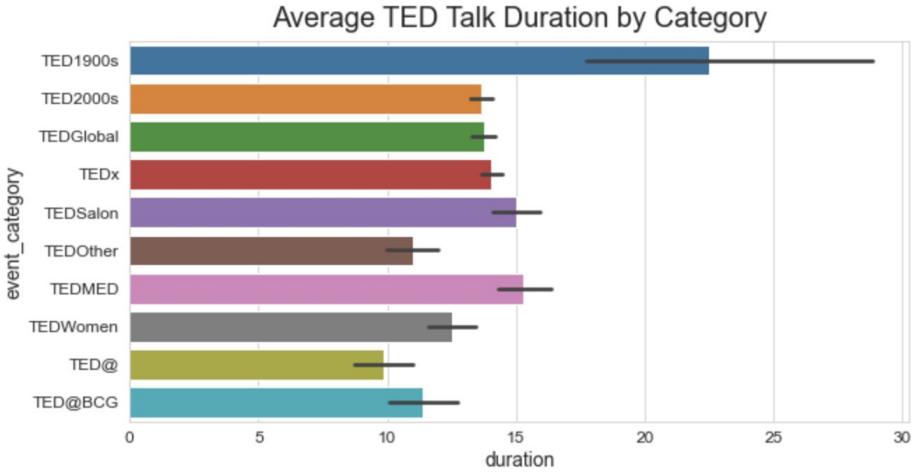


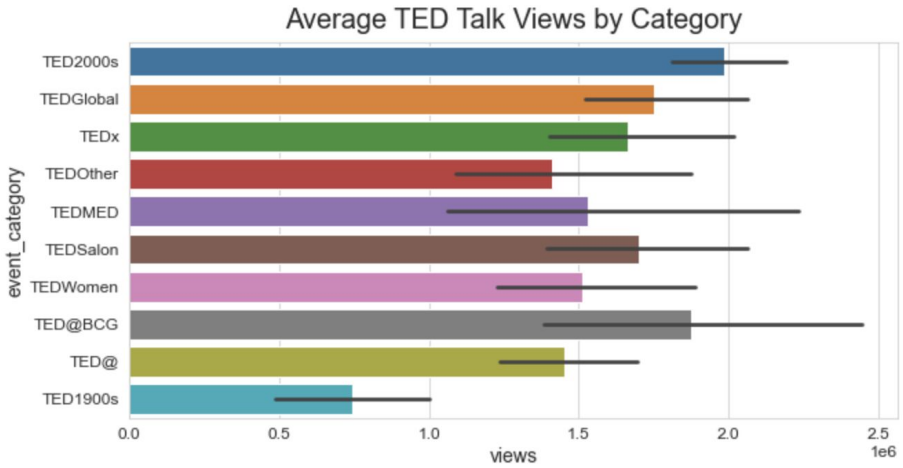
Methods - Part 2: EDA

1. Preliminary visualization of data
2. Check for correlation
3. Relationship between variables to gain insights



Methods - Part 2: EDA





The average TED talk was 13.6 mins long

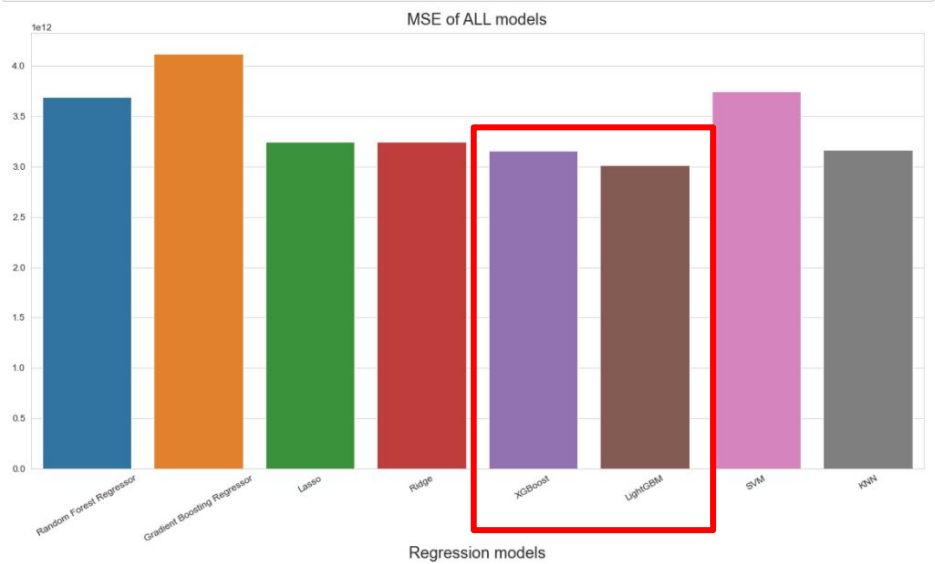
The average TED talk had 1.79 million views with a standard deviation of 2.71 million

Methods - Part 4:
Feature Engineering & Selection

1. **Removing Invalid features**
    - Drop views of related videos and duration of related videos, as we assume this will not be generated until after video is posted online.
    - Drop languages as TED uses a volunteer based translation service where viewers can translate their favourite talks after it is posted.
2. **New Features from Feature Engineering**
    - Repeat speaker
    - Length of title
    - Length of description
    - Binary topic categories (i.e. Tech, Art, Business)
3. **Feature selection**
    - Performing recursive feature elimination using random forest and selecting the top 35 features

Modeling

Various linear regression models were tested without any fine tuning to get a general idea of the performance given the current dataset and its selected features. All the models were trained using the second set of train data and their MSE and MAE were calculated using the second set of test data.

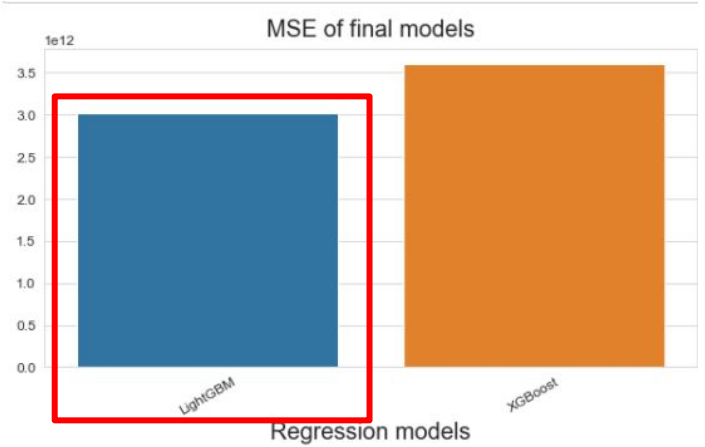


	Model	MSE	MAEs
5	LightGBM	3.018098e+12	8.678438e+05
4	XGBoost	3.158215e+12	9.456133e+05
7	KNN	3.167443e+12	9.351104e+05
3	Ridge	3.246171e+12	9.802539e+05
2	Lasso	3.250046e+12	9.816568e+05
0	Random Forest Regressor	3.694912e+12	1.056293e+06
6	SVM	3.750367e+12	9.199708e+05
1	Gradient Boosting Regressor	4.122359e+12	1.011968e+06

Modeling Results:
Top Models Performance

The top two performing models were fine tuned and their performance was evaluated using time-based cross validation.

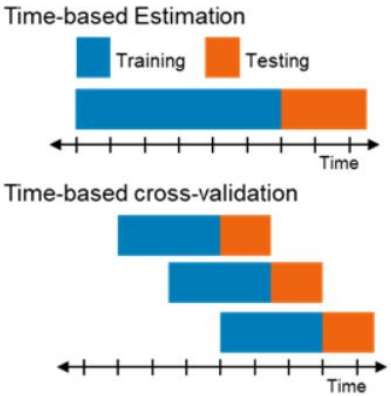
MODEL	MSE	MAE	% decrease in MSE
Light GBM	3 015 320 817 332.022 (3.015 * 10^12)	856 424.582	0.0920% decrease in MSE after tuning
XGBoost	3 609 111 569 540.2466 (3.609 * 10^12)	971 654.728	1.976% decrease in MSE after tuning



Modeling:
Time Based Cross Validation

Performance of Light GBM and XGBoost models was evaluated using time-based cross validation.

The time aspect and dimensionality of our data makes it relevant to evaluate it's performance using customized time-based cross validation solution that chooses relevant set sizes in terms of days.



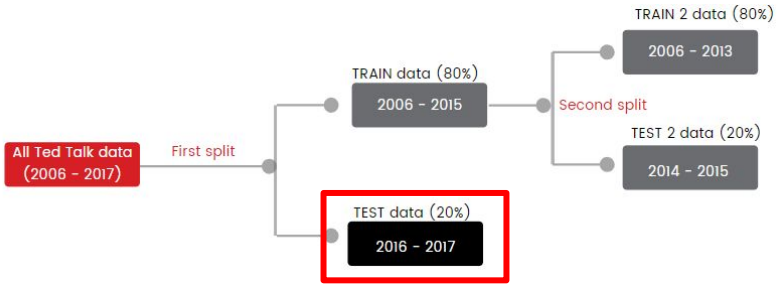

```
tscv = TimeBasedCV(train_period=7, test_period=3,freq='years')
index_output = tscv.split(data_for_modeling,
                          validation_split_date=datetime.date(2007,1,1))
```

Current split allows us to select relevant set sizes, as well as number of splits and also addresses the window size in terms of years

Ref: https://towardsdatascience.com/time-based-cross-validation-d259b13d42b8

Modeling Results:
Final model evaluation

Overall, LightGBM had the best performance after cross validation, thus in order to validate the model, we used the initial, untouched, test data to obtain the performance metrics.



Validation dataset for final model evaluation

Model	MAE	MSE	RMSLE
Tuned LGBM + test	**856 424.582**	3.015 *10 ^12	0.53472
Tuned LGBM + validation	**829 503.124**	2.211 * 10^12	0.41444

8 models tested → Light GBM, XGBoost → Fine tuning & cross validation → Fine tuned **LightGBM** model evaluation with validation aka “new data”

Modeling Results:
Feature importance

We performed this feature importance technique to evaluate what were the features that contributed the most to predicting the number of views for a TED talk.

Some main features used for prediction included:

1. Year
2. Day filmed
3. Month filmed
4. Month posted
5. Number of speakers
6. Speaker frequency
7. Category:
    a. Communication
    b. Technology/Science
    c. Humanity
    d. Global Issues
    e. Business

The most important features were:

1. **Duration**
2. **Description length**
3. **Title Length**

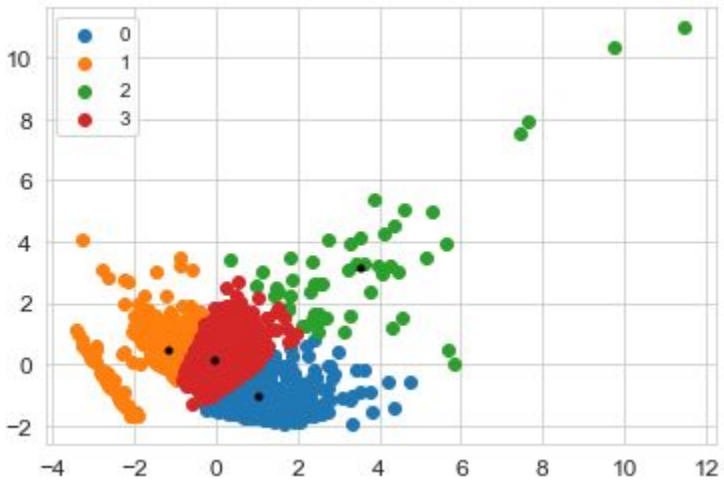
Clustering

Talk duration

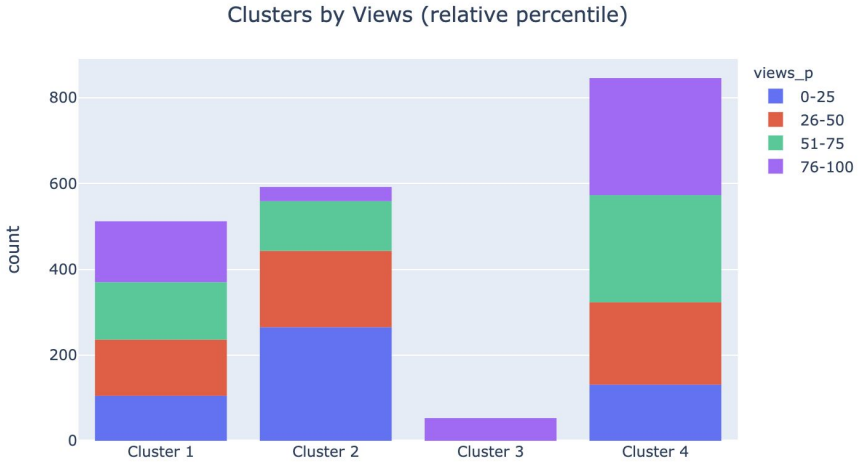
Views

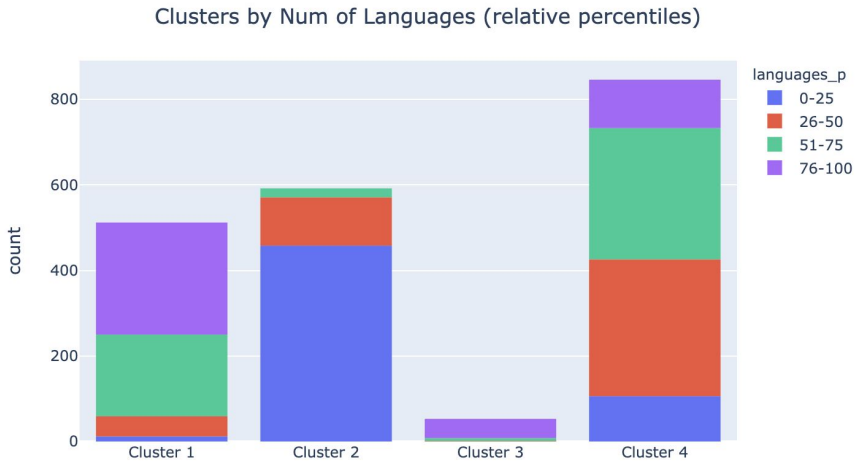
Num Languages

PCA Representation of Clusters



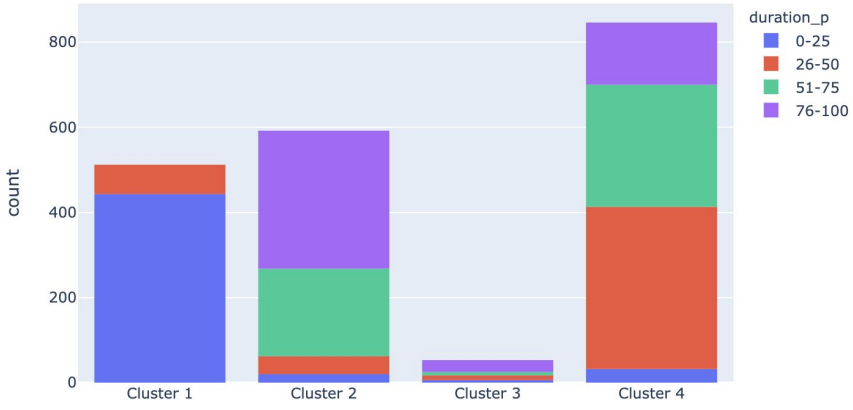
Clustering (cont.)





Cluster 1: shortest duration and most languages

Cluster 2: longest duration and least languages



Cluster 3: highest views, languages, and longest talks, but smallest group

Cluster 4: majority talks = mix of characteristics

Business Implications

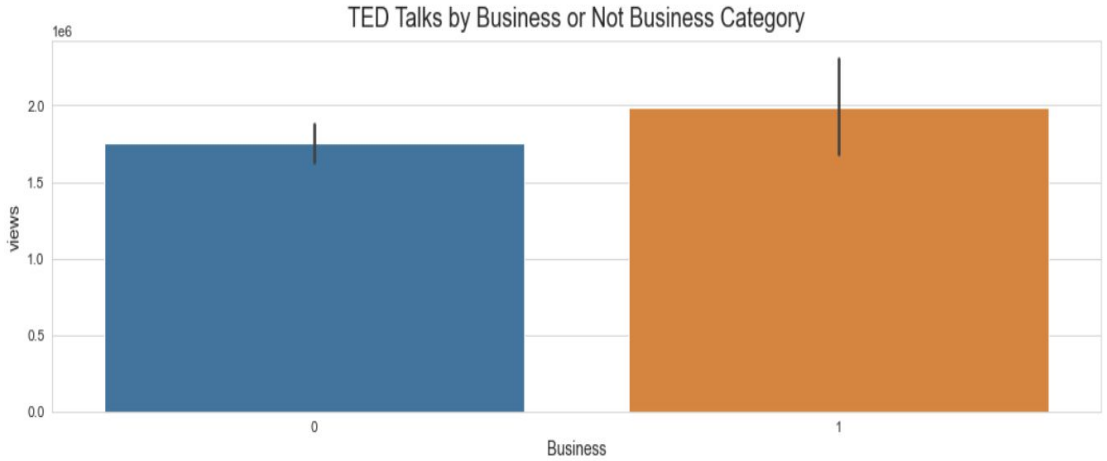
Features most important for prediction are immediately observable when browsing TED talks, as opposed to “hidden” features such as year filmed

Users may be easily influenced to view or not simply based on online visual cues -- by understanding these attributes, TED can thus easily modify and leverage these features to attract more viewers

The most important features were:

1. **Duration**
2. **Description length**
3. **Title Length**

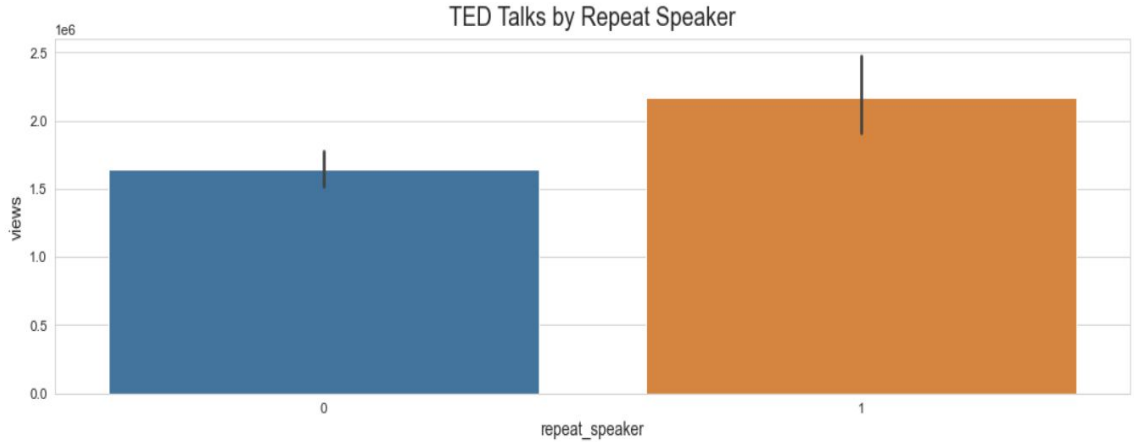
Characteristics of most viewed TED talks



Talks that were viewed more were more predominantly....

- Business related talks

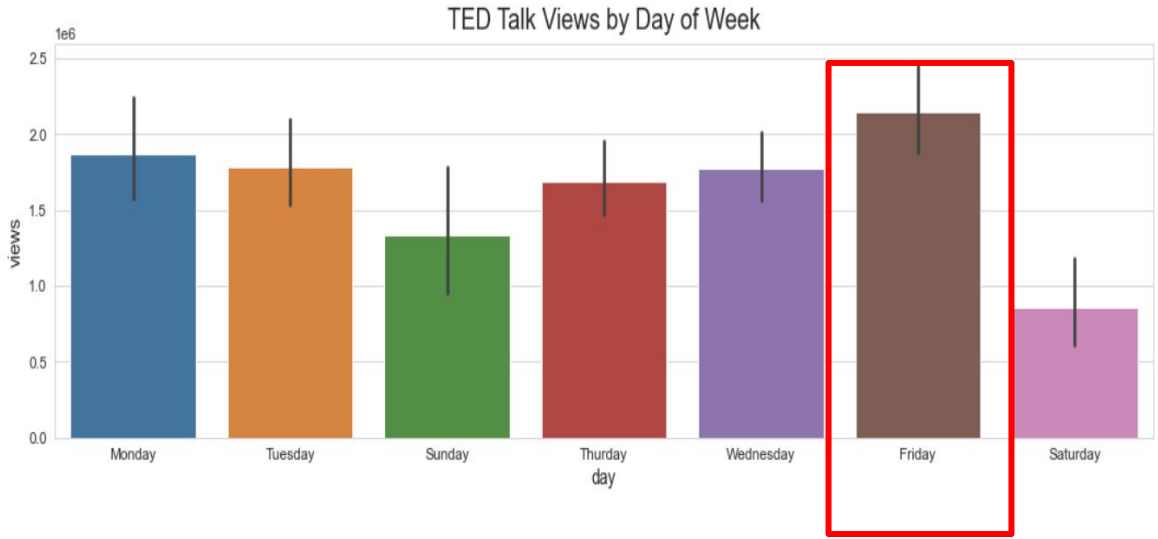
Characteristics of most viewed TED talks



Talks that were viewed more were more predominantly....

- Business related talks
- Talks featuring repeat speakers

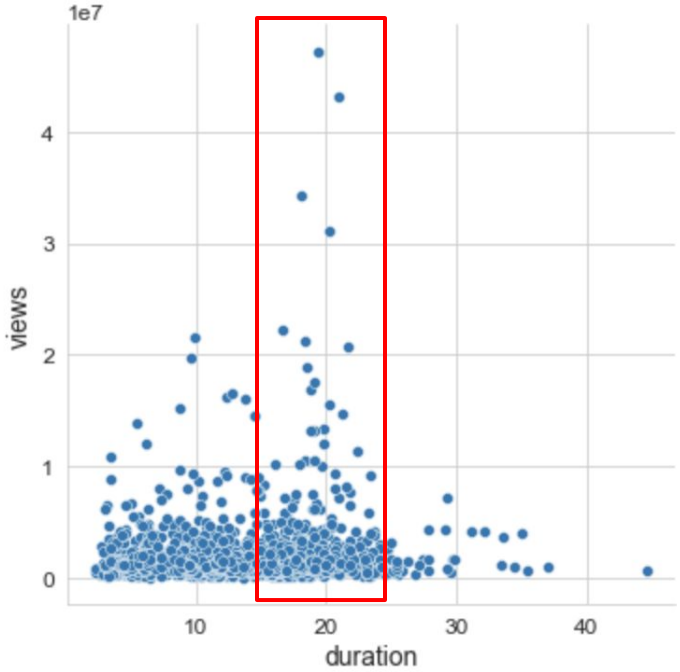
Characteristics of most viewed TED talks



Talks that were viewed more were more predominantly....

- Business related talks
- Talks featuring repeat speakers
- Published on Friday

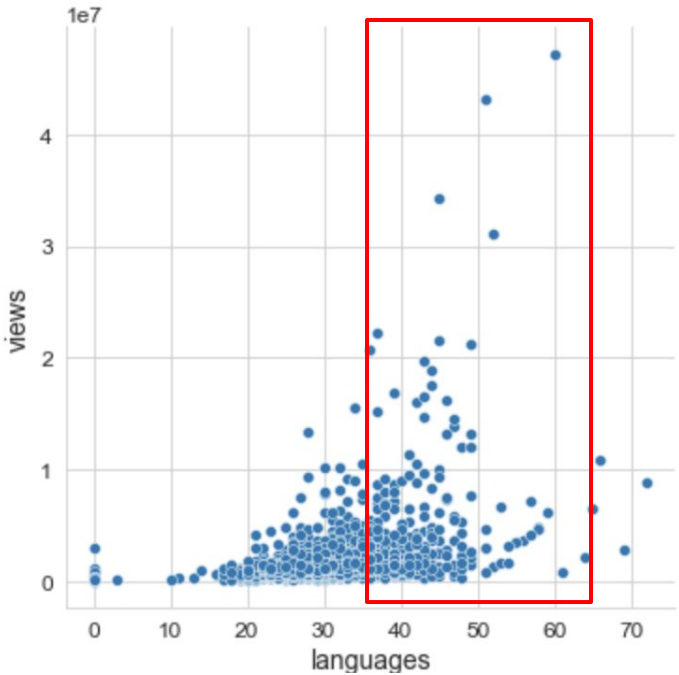
Characteristics of most viewed TED talks



Talks that were viewed more were more predominantly....

- Business related talks
- Talks featuring repeat speakers
- Published on Friday
- Around 20 minutes

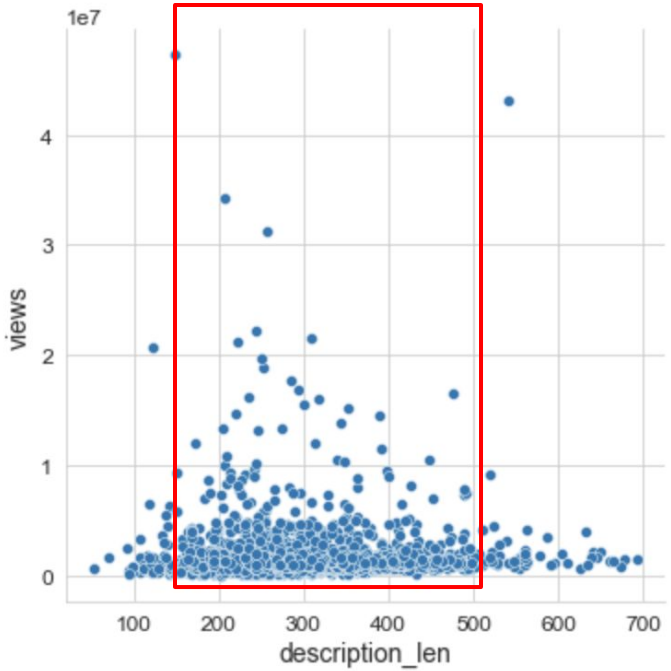
Characteristics of most viewed TED talks



Talks that were viewed more were more predominantly....

- Business related talks
- Talks featuring repeat speakers
- Published on Friday
- Around 20 minutes
- Translated into many languages

Characteristics of most viewed TED talks



Talks that were viewed more were more predominantly....

- Business related talks
- Talks featuring repeat speakers
- Published on Friday
- Around 20 minutes
- Translated into many languages
- Medium description lengths

Discussion and Limitations

Methodological Limitations

- Feature importance does not show direction nor the causality of the relationship

Data Limitations

- Factors that influences popularity but is not represented by the data (geographical location of the talk, advertising campaign of a given event, etc)
- *Potential overlay data*: Social media trends, engagements on social media (Tweets, Google searches etc)

Future Steps

- *Causal inference*: determining the factors (causes) that determine views (effect)
- Try different targets or tasks (classification task)
- Using insights from clustering/segmentations to further “group” Ted Talks before prediction for business and marketing purposes

Conclusion

Hypothesis testing:

While Gradient Boosting Regressor did not perform most optimally, our hypothesis was **supported** since tree-based methods, specifically LightGBM and XGBoost, outperformed the other models at prediction. However, this difference is quite marginal.

We ***cannot conclude that the current model is effective*** and performant at predicting the number of views of future Ted Talks, based on the metrics evaluated (MAE, MSE, RMSLE).

Contrary to what we expected, features such as ***duration, description length, and title length were most important for predicting the number of views***, as opposed to features more directly related to the content of the talk itself, such as its topic category.

TED Ideas worth spreading

Thanks

for coming to our TED Talk