TED Talk Performance Predictors

-An application example of

Random Forest
Natural Language Processing
Network Analytics
Negative Binomial Regression

Xiaoyan Zhou
Data Analytics Intern in GwIA

CONTENT

1. Introduction
2. Data Description and Analysis
3. Variable Generation and Evaluation
4. Model Building and Results
5. Managerial Recommendations
6. Limitation and Future Research

TED

1. Intro

Objectives:

Find out the important variables affecting the views of TED Talks

- TED Talks producers can better organize their talks
- TED.com can improve the demonstration of the talks by optimizing relevant attributions.

Toolkit:

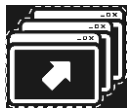
Python Jupyter Notebook/R

1. Intro

Executive Summary:

TED Talk Speakers:

- The theme of TED Talks can be adjusted to cater for the flavor of the audiences.
- Improve the style of talks and descriptions



TED.com:

- Reorganize the recommendation list and expand the list.
- Promote TEDx event
- Encourage more female TED Talk speaker to join

2. Data Description& Analysis

2.1 Data Description

- Ted_main.csv
- Trainscript.csv: transcripts for all talks.

Observation: 2247.

Data source :
https://www.kaggle.com/rounakbanik/ted-talks

Columns	Descriptions
comments	The number of comments of the video.
description	A brief introduction of what the talk is about.
duration	The total seconds of the talk.
event	The TED/TEDx event where the talk took place.
film_date	The Unix timestamp of the filming.
languages	The number of languages that the talks is available.
main_speaker	The first name speaker of the talk.
num_speaker	The number of speakers in the talk.
publish_date	The Unix timestamp when the talk was published in TED.com
ratings	Ratings given to the talk, including the name of the ratings (Funny, Beautiful, Obnoxious, etc.) and the count of each sort of rating.
related_talks	Recommended talks to watch next.
speaker_occupation	The occupation of the first name speaker.
tags	The themes associated to the talk.
title	The title of the talk.
urls	The URL of the talk.
views	The number of views on the talk.

2. Data Description& Analysis

2.2 Data Analysis

(1) The number of TED Talks and views in each year: both surge quickly after year 2008

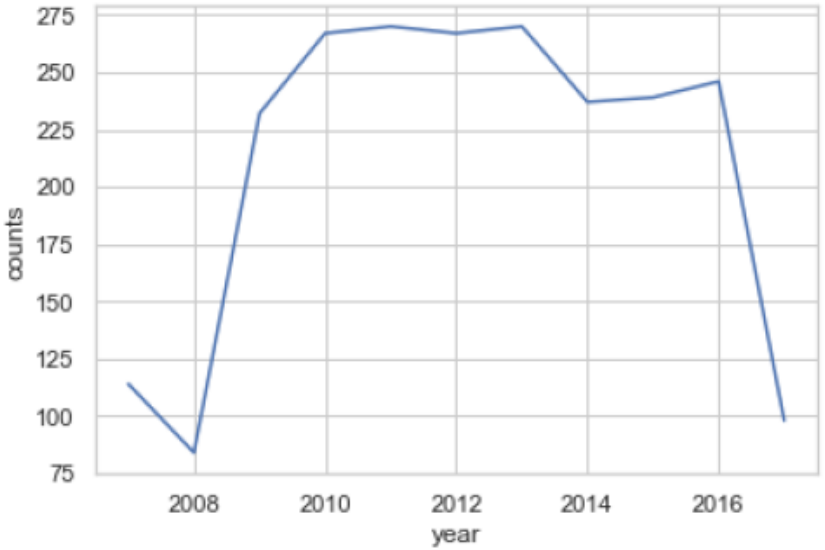


Figure 2.2.3 TED Talks counts by year

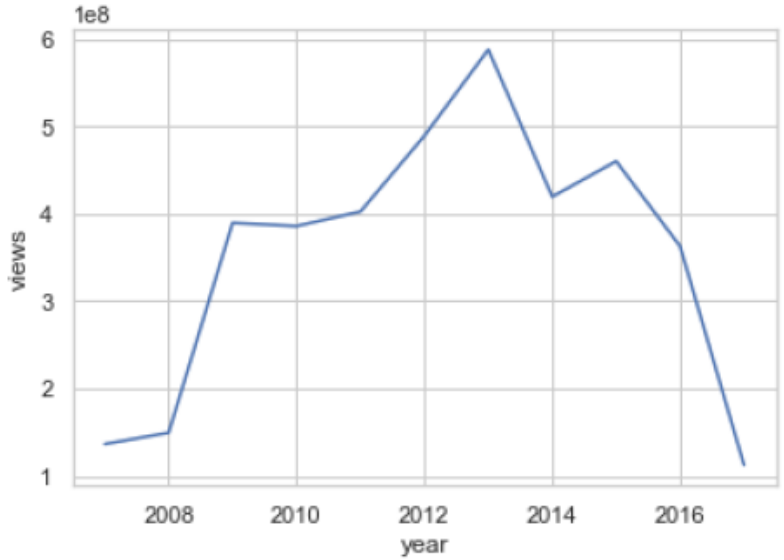


Figure 2.2.4 TED Talks views by year

Before 2009, the number of TED Talks released each year are relatively small, while after 2009, the number of TED Talks released each year are above 225.

TED Talks views increase quickly after the year 2008, peaking in the year 2013, and start to decrease afterward.

2. Data Description& Analysis

2.2 Data Analysis

(2) Popular themes among TED Talk producer and audience are different

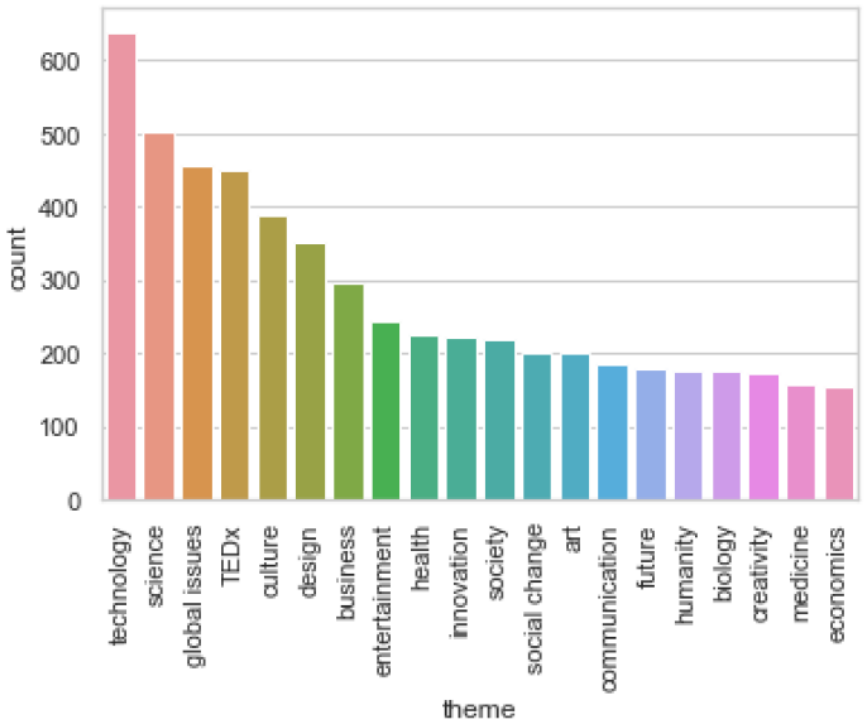


Figure 2.2.1 Theme Popularity- Count

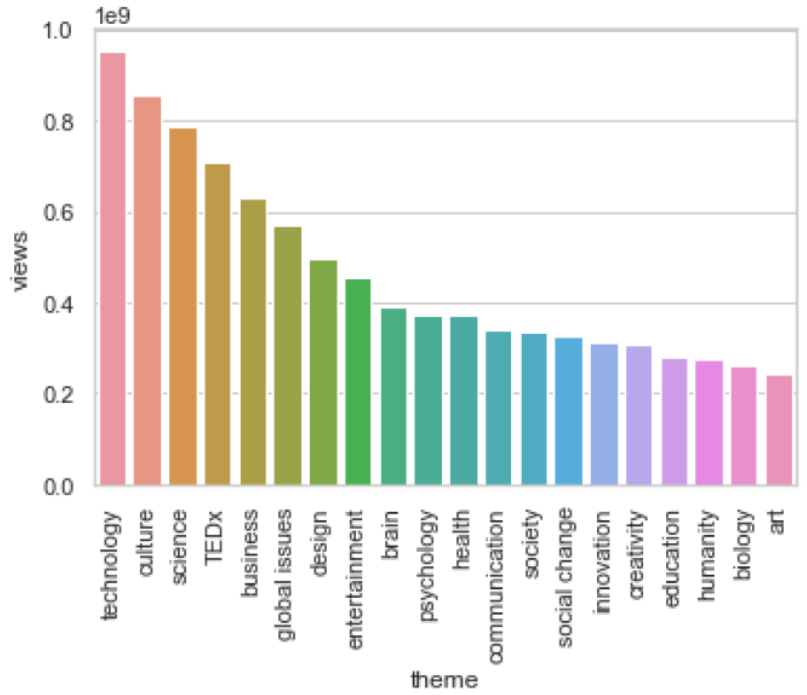


Figure 2.2.2 Theme Popularity- Views

Count:
Future
Medicine
Economics

Views:
Brain
Psychology
Social change

2. Data Description& Analysis

2.2 Data Analysis

(3) Word cloud of title: different aspects of concerns.

Figure 2.2.5 World Cloud for all TED Talks Title

Figure 2.2.6 World Cloud for 100 Most Viewed TED Talks Title

All talks:
The TED Talks speakers tend to concern about the influencing the life or the world.

Most views:
More interpersonal, more about improving daily life or work.

3. Variable Generation& Evaluation

3.1 Variable Generation

Method: NLP(Natural Language Processing)

Data Require:
Text data, ie: email, transcript, name, address.

Application:
Style/Theme Analytics

- Description Positivity

Python NLTK package

Data example

Sir Ken Robinson makes an entertaining and profoundly moving case for creating an education system that nurtures (rather than undermines) creativity.

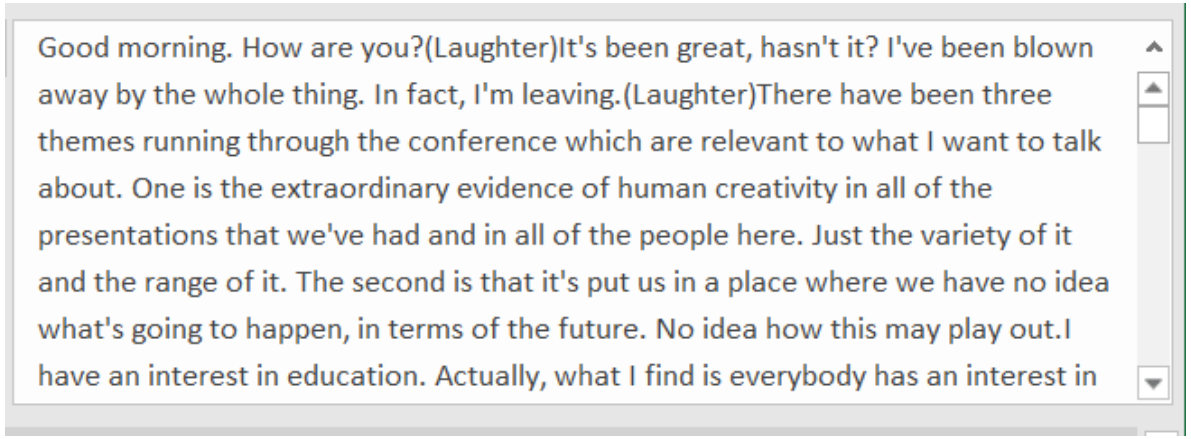
- Main Speaker Gender

Python gender_guesser package

Ken Robinson
Julia Sweeney

- Incitement

How many times the speaker triggers the audiences to laugh or applause during the talk



3. Variable Generation& Evaluation

3.1 Variable Generation

Method: Network Analytics

Data required: Data that measures the relationship between objects: social network, communication patterns

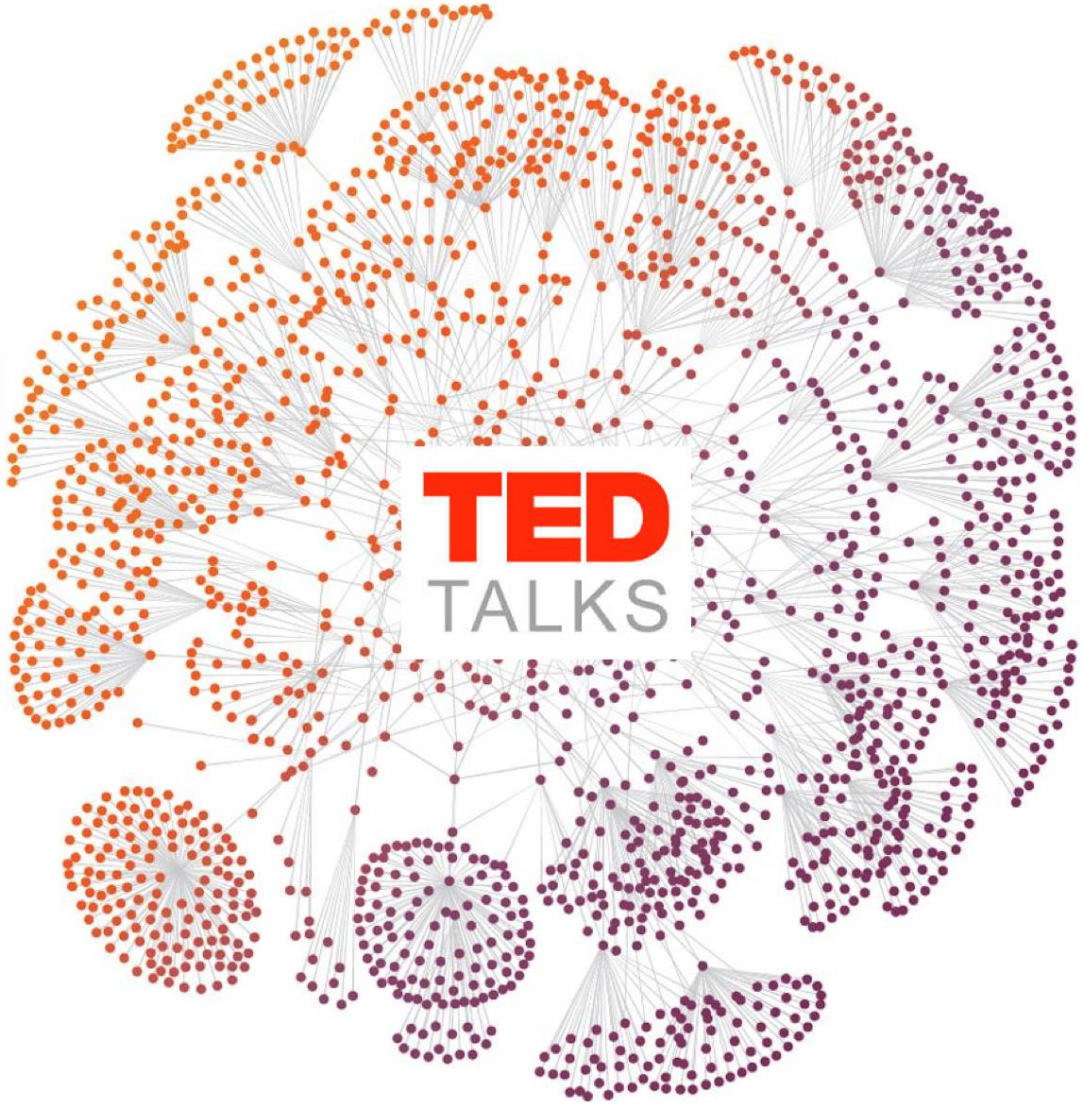
Application: Identify influencer in a network; Segmentation/Community detection.

- Indegree

How many other talks recommend this talk to play next

- Eigenvector Centrality

A TED Talk with high eigenvector centrality means that it is recommended by many other TED Talks who are recommended by many other TED Talks



3. Variable Generation& Evaluation

3.1 Variable Generation

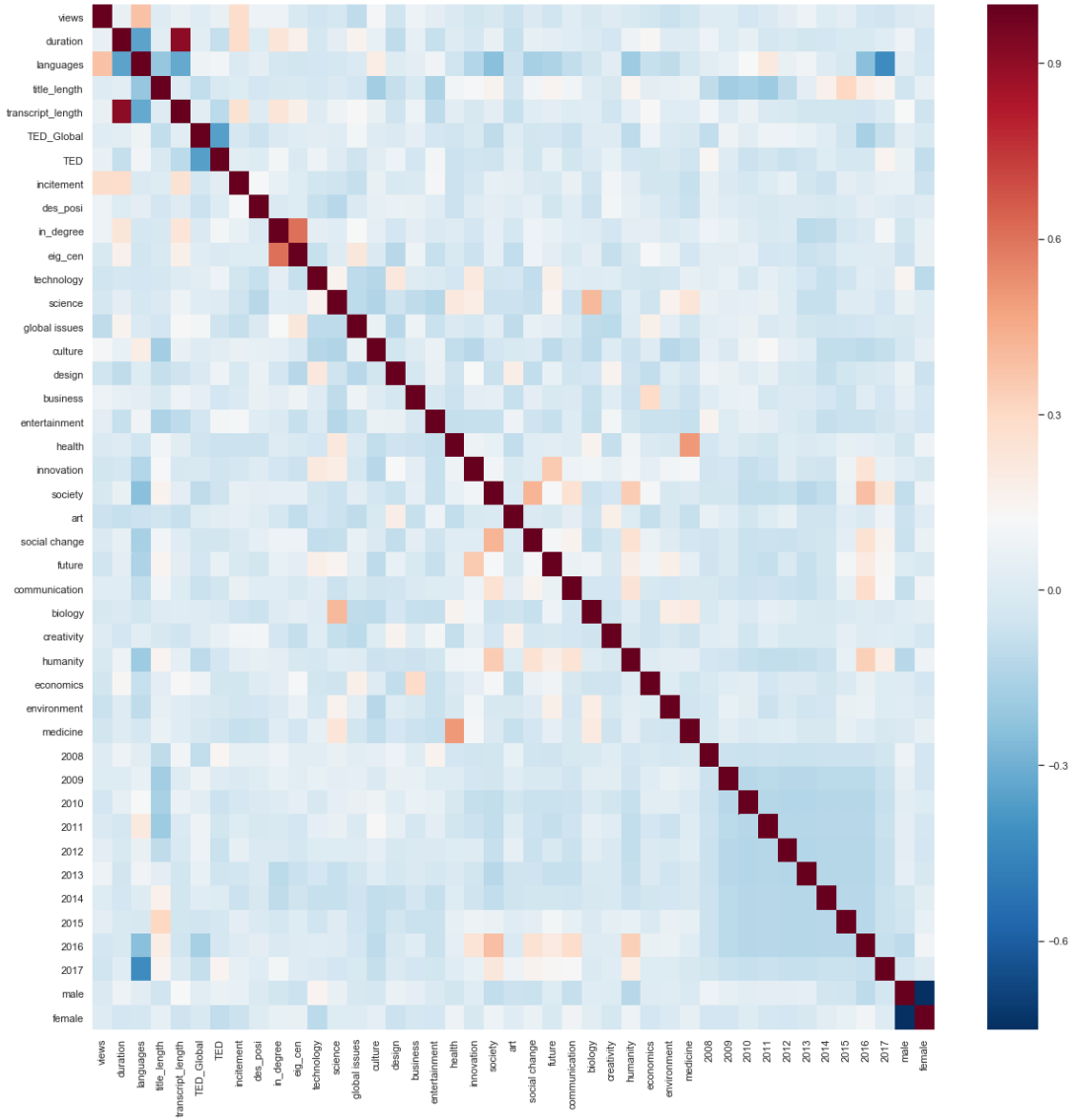
- Duration
- Year Dummy
- TED Global/TED Dummy
- 20 Themes Dummy

3. Variable Generation& Evaluation

3.1 Variable Evaluation

Step 1: Detecting Sources of Multicollinearity, use correlatio heatmap to help selecting variables.

- Exclude variable with large VIF values, ie: transcript positiveness, number of speaker, PageRank centrality of TED Talks.
- VIF under 10 is acceptable.
- Full VIF is in Appendix 1



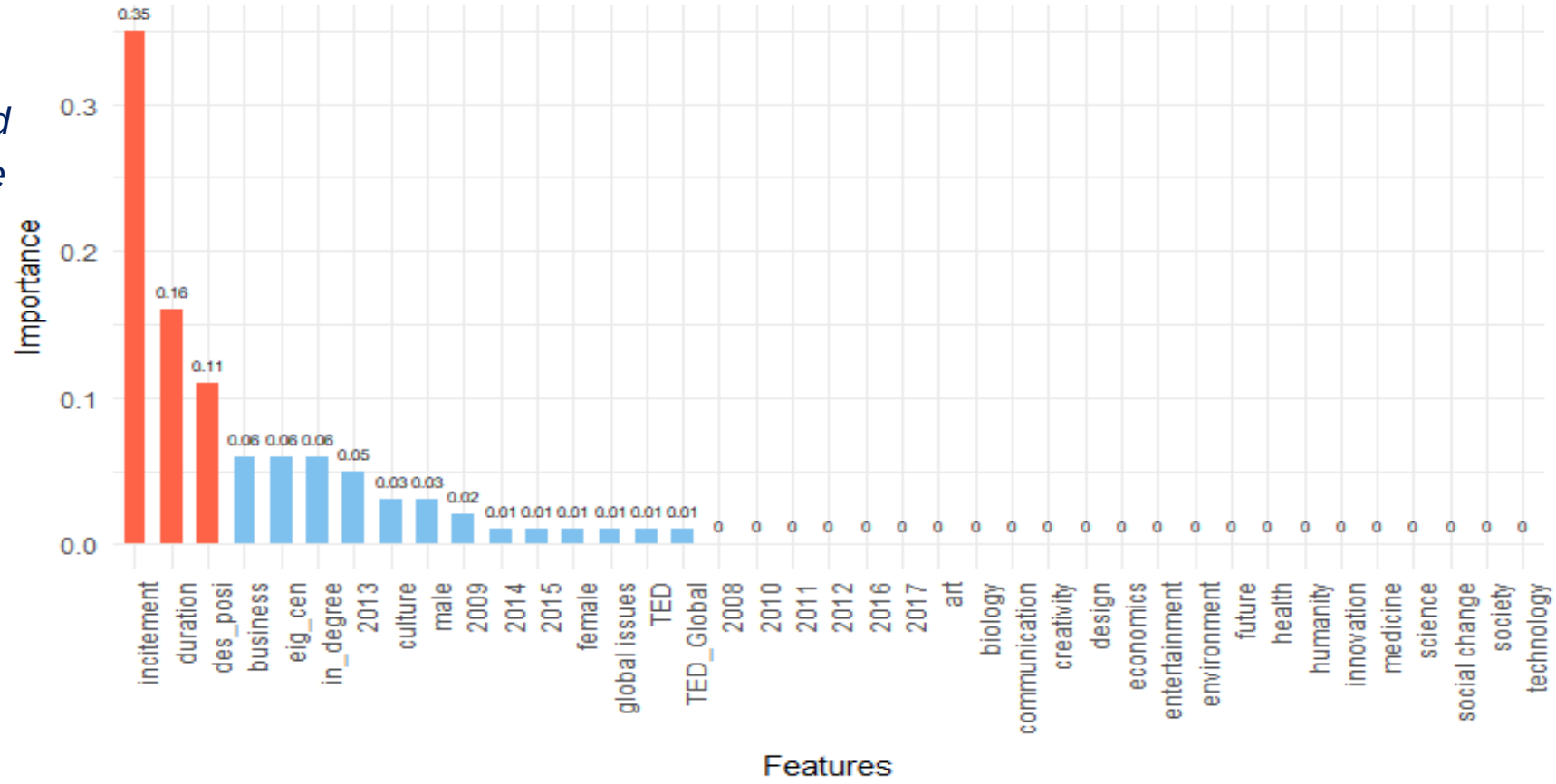
3. Variable Generation& Evaluation

3.1 Variable Evaluation

Step 2: Feature importance- Random Forest Algorithm (Machine learning methods)

Data required: Targeted factors and relative features that may affect the target

Application: Feature Selection/ Predictive Analytics: Regression& Classification



4. Model Building and Results

Objective:

Find out how the unit change of dependent variables can influence the unit change of independent variables.

Model: Negative Binomial regression

- Dependent variable is count data
- Heavily right-skewed and over dispersed
- Way higher variance (5469585245594) compared to the mean (1722348).

4. Model Building and Results

Results:

What matters:

- The event of the TED Talks
- Description positiveness
- Main speaker gender: female speakers out performance
- Theme(Popular: culture, business, and health; Unpopular: global issues, art, environment, and medicine)
- Network attributes: Indegree; eigenvector centrality
- Year

What doesn't matter:

- Duration
- Some other theme: technology, science

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	13.1507	0.147	89.605	0.000	12.863	13.438
TED_Global	0.1586	0.067	2.354	0.019	0.027	0.291
TED	0.2145	0.057	3.774	0.000	0.103	0.326
incitement	0.0362	0.004	9.519	0.000	0.029	0.044
des_posi	1.3868	0.514	2.698	0.007	0.379	2.394
in_degree	0.0432	0.008	5.121	0.000	0.027	0.060
eig_cen	-3.8754	1.668	-2.323	0.020	-7.145	-0.605
global_issues	-0.3405	0.064	-5.357	0.000	-0.465	-0.216
culture	0.2198	0.067	3.262	0.001	0.088	0.352
business	0.3403	0.073	4.687	0.000	0.198	0.483
health	0.2129	0.092	2.316	0.021	0.033	0.393
art	-0.2429	0.089	-2.719	0.007	-0.418	-0.068
environment	-0.2029	0.098	-2.065	0.039	-0.396	-0.010
medicine	-0.3047	0.107	-2.855	0.004	-0.514	-0.096
year_2008	0.1881	0.169	1.112	0.266	-0.144	0.520
year_2009	0.2707	0.133	2.033	0.042	0.010	0.532
year_2010	0.2069	0.133	1.561	0.118	-0.053	0.467
year_2011	0.3295	0.131	2.508	0.012	0.072	0.587
year_2012	0.4707	0.131	3.602	0.000	0.215	0.727
year_2013	0.7122	0.131	5.454	0.000	0.456	0.968
year_2014	0.5824	0.134	4.345	0.000	0.320	0.845
year_2015	0.6145	0.132	4.672	0.000	0.357	0.872
year_2016	0.3438	0.134	2.567	0.010	0.081	0.606
year_2017	-0.0050	0.162	-0.031	0.975	-0.323	0.313
male	0.1326	0.074	1.804	0.071	-0.011	0.277
female	0.1586	0.080	1.980	0.048	0.002	0.316

Figure 5.2.1 Summary for the final Negative Binomial Model

5. Managerial Recommendation

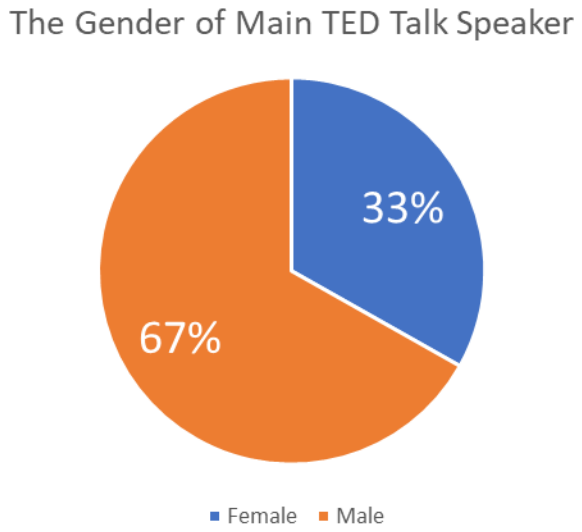
TED Talk Speaker:

- Consider the topics that are favored by the audiences or make the talk more relevant to those popular topics.
- Articulate the expression, causing more laugh or applause during the speech. Use more positive words to describe the talk.

Medicine → Health

TED.com:

- Encourage and introduce more female speaker to present the talks
- Expand the watch next list.



Both the producer and the TED.com can use the negative binomial model trained in this report to predict views for a talk before publishing it

6. Limitation and Future Research

Limitations:

- Gender variables generated in this report are done by the gender guesser algorithm in Python, not as accurate as collecting the information directly.
- The views of TED Talks is considered as a proxy of the TED Talks performances, cannot reflect other aspect of performance – can be improved by using “effectiveness”, “rating positiveness”

Future research:

- The eigenvector centrality has a negative influence on the TED Talks views, this result is counterintuitive. Further research is needed to figure out the causal relationship.

(The reason behind this might be that TED.com tend to recommend those unpopular TED Talks after one TED Talk is finished playing.)

- K nearest neighboring methods can be used, as this model can match similar TED Talks and predict for the views according to their attributes.

Q&A

Thank you for listening.

Xiaoyan Zhou

Appendix 1: VIF for Variables Created

	features	VIF Factor
0	duration	8.090003
1	TED_Global	1.620483
2	TED	1.908517
3	incitement	2.228773
4	des_posi	3.310554
5	in_degree	5.342747
6	eig_cen	2.476481
7	technology	1.687148
8	science	1.850257
9	global issues	1.471718
10	culture	1.362676
11	design	1.403145
12	business	1.310569
13	entertainment	1.230491
14	health	1.559355
15	innovation	1.470244
16	society	1.705853
17	art	1.215179
18	social change	1.421651
19	future	1.410883
20	communication	1.292467
21	biology	1.378306
22	creativity	1.192926
23	humanity	1.436990
24	economics	1.274833
25	environment	1.214830
26	medicine	1.530855
27	2008	1.496490
28	2009	2.011585
29	2010	2.143481
30	2011	2.019762
31	2012	2.096947
32	2013	2.064916
33	2014	1.904096
34	2015	2.101177
35	2016	2.693034
36	2017	1.721869
37	male	5.170769
38	female	3.045922

Appendix 2: Summary of the Full Model

Model:	GLM	Df Residuals:	1632
Model Family:	NegativeBinomial	Df Model:	39
Link Function:	log	Scale:	0.832915921584
Method:	IRLS	Log-Likelihood:	-23977.
Date:	Sun, 20 Aug 2018	Deviance:	949.29
Time:	17:56:08	Pearson chi2:	1.28e+03
No. Iterations:	22		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	13.1972	0.161	81.773	0.000	12.881	13.514
duration	-1.913e-05	8.11e-05	-0.024	0.981	-0.000	0.000
TED_Global	0.1827	0.066	2.765	0.006	0.054	0.311
TED	0.2184	0.056	3.885	0.000	0.108	0.329
incitement	0.0350	0.004	8.984	0.000	0.027	0.043
des_posi	1.3519	0.501	2.698	0.007	0.369	2.335
in_degree	0.0435	0.008	5.195	0.000	0.027	0.060
eig_cen	-3.8285	1.650	-2.321	0.020	-7.062	-0.595
technology	-0.0841	0.057	-1.470	0.142	-0.196	0.028
science	0.0879	0.068	1.288	0.198	-0.046	0.222
global_issues	-0.3309	0.064	-5.149	0.000	-0.457	-0.205
culture	0.2080	0.067	3.126	0.002	0.078	0.338
design	-0.1122	0.072	-1.570	0.117	-0.252	0.028
business	0.3963	0.079	5.000	0.000	0.223	0.510
entertainment	0.0359	0.083	0.434	0.664	-0.126	0.198
health	0.1944	0.090	2.159	0.031	0.018	0.371
innovation	-0.0441	0.090	-0.492	0.023	-0.220	0.132
society	-0.0567	0.098	-0.578	0.563	-0.249	0.136
art	-0.2352	0.089	-2.649	0.008	-0.409	-0.061
social_change	-0.1291	0.093	-1.383	0.163	-0.311	0.052
future	-0.0040	0.099	-0.040	0.968	-0.197	0.189
communication	0.0350	0.093	0.376	0.707	-0.147	0.217
biology	-0.1791	0.097	-1.846	0.065	-0.369	0.011
creativity	-0.0056	0.094	-0.060	0.952	-0.190	0.179
humanity	0.1222	0.100	1.225	0.221	-0.073	0.318
economics	-0.1968	0.100	-1.961	0.050	-0.393	-0.000
environment	-0.1909	0.099	-1.993	0.046	-0.391	-0.003
medicine	-0.2908	0.105	-2.773	0.006	-0.496	-0.085
year_2008	0.1761	0.166	1.069	0.285	-0.147	0.499
year_2009	0.2644	0.131	2.016	0.044	0.007	0.521
year_2010	0.1907	0.130	1.464	0.143	-0.064	0.444
year_2011	0.3364	0.130	2.582	0.010	0.082	0.591
year_2012	0.4576	0.130	3.508	0.000	0.202	0.713
year_2013	0.6933	0.130	5.330	0.000	0.438	0.948
year_2014	0.5539	0.133	4.169	0.000	0.293	0.814
year_2015	0.6078	0.134	4.529	0.000	0.345	0.871
year_2016	0.3737	0.145	2.582	0.010	0.090	0.657
year_2017	0.0168	0.169	0.099	0.921	-0.315	0.349
male	0.1464	0.072	2.031	0.042	0.005	0.288
female	0.1549	0.078	1.980	0.048	0.002	0.308