Predict the popularity of a TED Talk

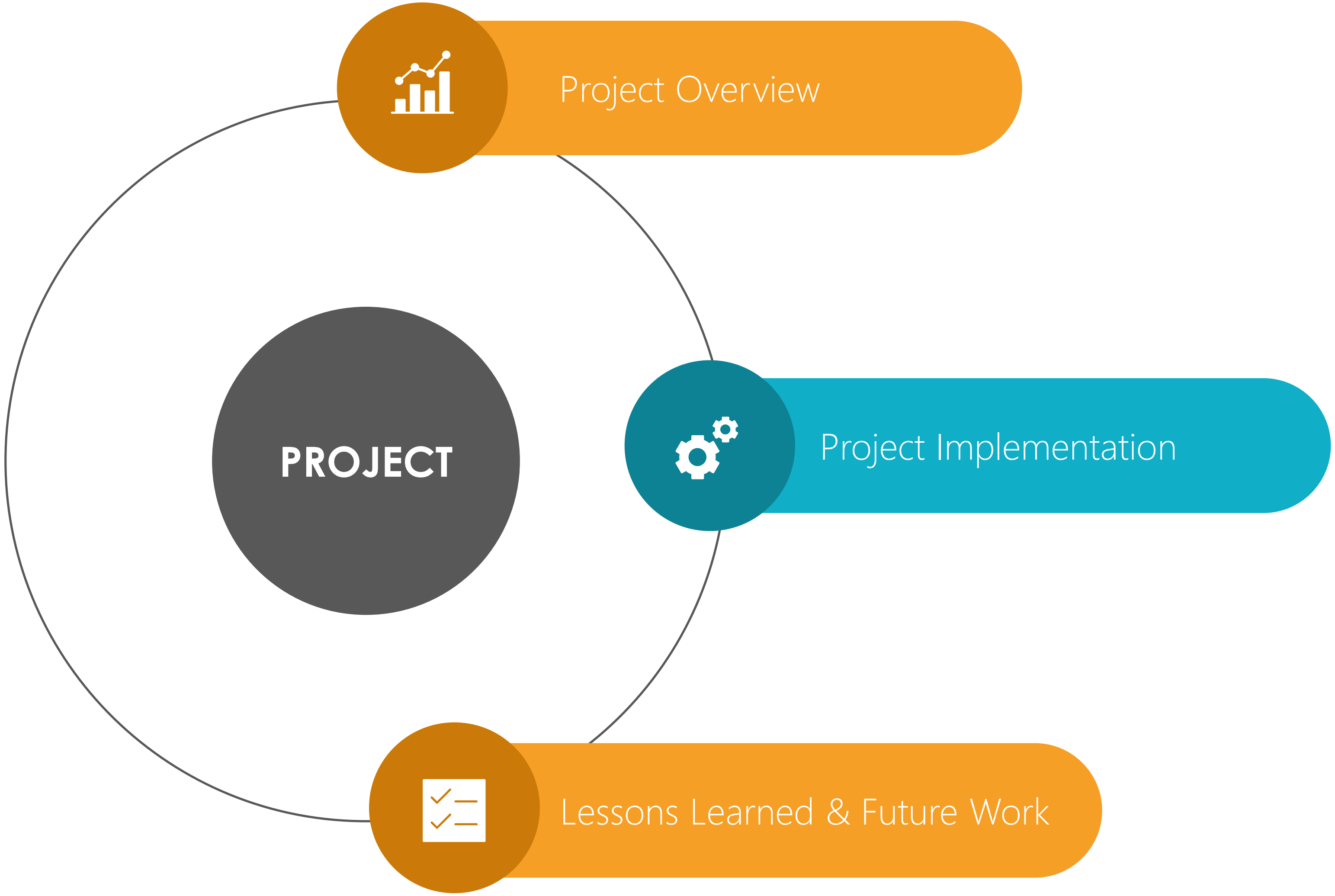
CS522 Course Project

Name: Liu, Shijie
Professor: Liu, Xinlian
Date: 12/10/2018

Computer Science Department
Hood College of Frederick Maryland

TED Ideas worth spreading

Project Outline



PROJECT

- Project Overview
- Project Implementation
- Lessons Learned & Future Work

Project Overview

Problem:

- TED Conferences LLC is a media organization that posts talks online for free distribution under the slogan "ideas worth spreading." People could also organize their own local TEDx event.
- How could TED filter which talk shall be published?
- What could an organizer do to make their talk more popular?

Objectives:

- What makes a popular TED Talk?
- What are the most popular Topics that people like talk about?
- Predict the popularity of an un-published TED Talk based on transcript.

Implementation

Data Collection

The Datasets being used for this project is being downloaded from Kaggle.

Data Understanding

Using Exploratory data analysis(EDA) Technique to get a better understanding on the characteristics for existing attributes.

Data Pre-Processing

Handle missing values if there is any, normalize attributes upon needed, and add new attributes upon necessary.

Model Building

Build various Classification Models using different Algorithms and train the model with training dataset.

Model Selection

Compare the performance and accuracy of different models and select the best Classification Model.

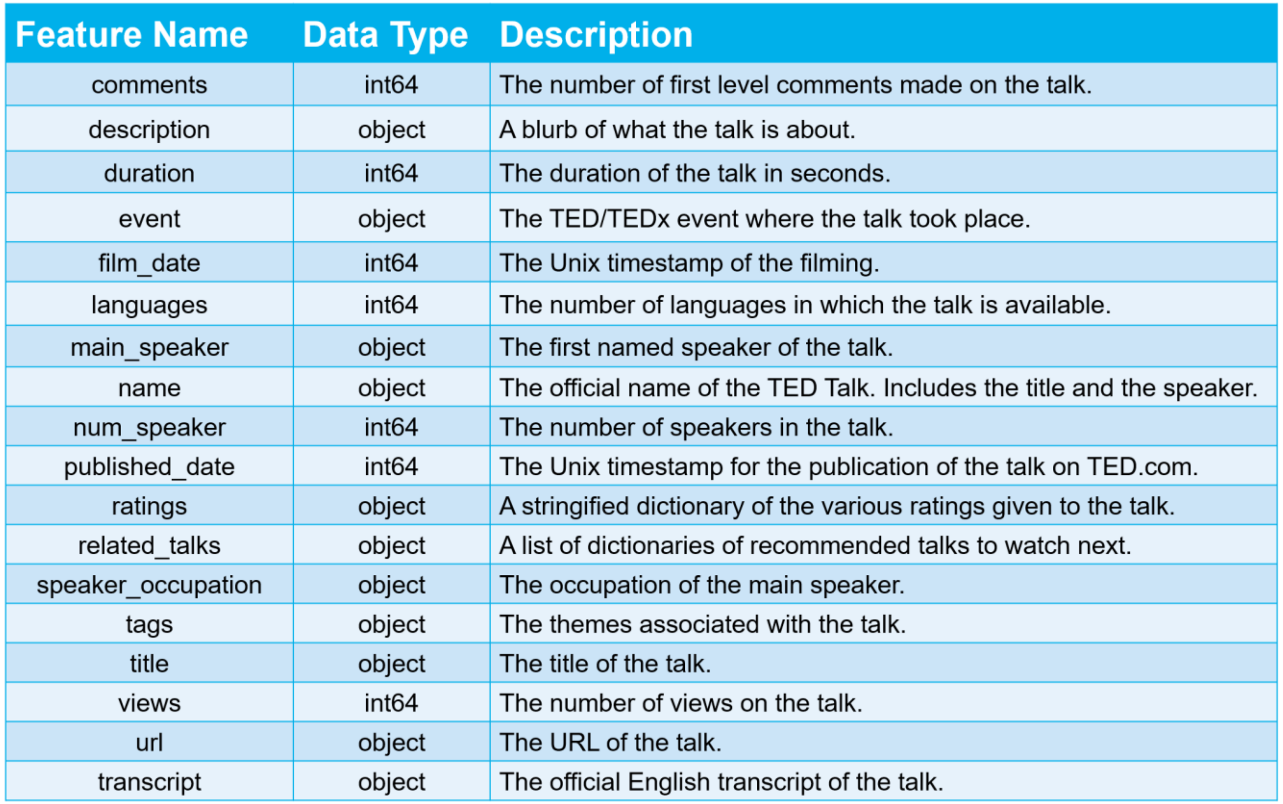
Data Collection

Data Source

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

TED Main
(2550 x 17)

TED Transcript
(2467 x 2)



Feature Name	Data Type	Description
comments	int64	The number of first level comments made on the talk.
description	object	A blurb of what the talk is about.
duration	int64	The duration of the talk in seconds.
event	object	The TED/TEDx event where the talk took place.
film_date	int64	The Unix timestamp of the filming.
languages	int64	The number of languages in which the talk is available.
main_speaker	object	The first named speaker of the talk.
name	object	The official name of the TED Talk. Includes the title and the speaker.
num_speaker	int64	The number of speakers in the talk.
published_date	int64	The Unix timestamp for the publication of the talk on TED.com.
ratings	object	A stringified dictionary of the various ratings given to the talk.
related_talks	object	A list of dictionaries of recommended talks to watch next.
speaker_occupation	object	The occupation of the main speaker.
tags	object	The themes associated with the talk.
title	object	The title of the talk.
views	int64	The number of views on the talk.
url	object	The URL of the talk.
transcript	object	The official English transcript of the talk.

Data Collection

TED Main- Example Data:

comments	description	duration	event	film_date	languages	main_speaker	name	num_speaker	published_date	ratings	related_talks	speaker_occupation	tags	title	url	views
4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	1140825600	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 19645}, {...	[{'id': 865, 'hero': 'https://pe.tedcdn.com/im...	Author/educator	['children', 'creativity', 'culture', 'dance', ...	Do schools kill creativity?	https://www.ted.com/talks/ken_robinson_says_sc...	47227110

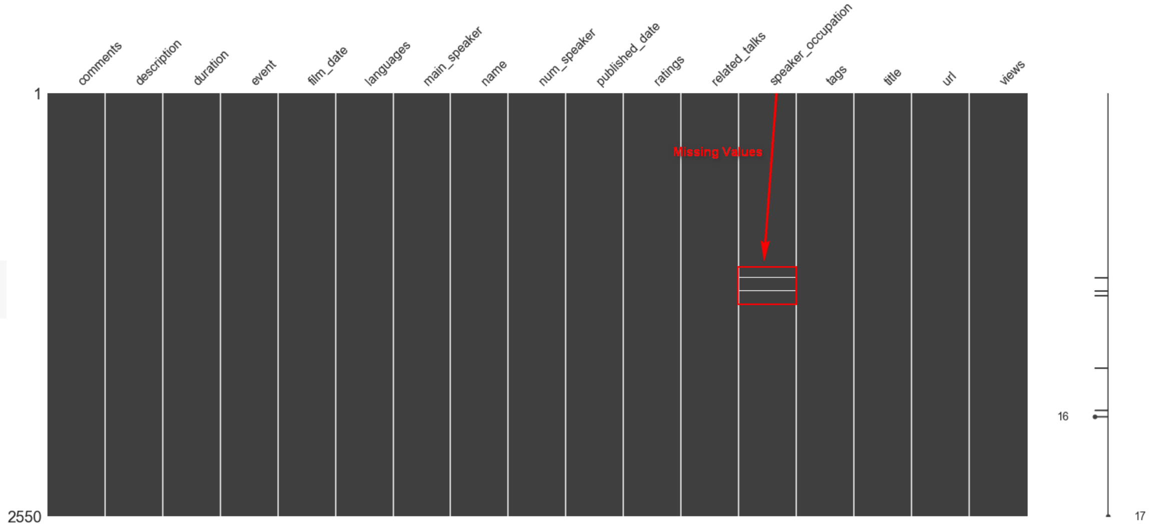
TED Transcript -Example Data:

transcript	url
Good morning. How are you?(Laughter)It's been great, hasn't it? I've been blown away by the whole thing. In fact, I'm leaving.(Laughter)There have been three themes running through the conference which are relevant to what I want to talk about. One is the extraordinary evidence of human creativity in all of the presentations that we've had and in all of the people here. Just the variety of it and the range of it. The second is that it's put us in a place where we have no idea what's going to happen, in terms of the future. No idea how this may play out.I have an interest in education. Actually, what I find is everybody has an interest in education. Don't you? I find this very interesting. If you're at a dinner party, and you say you work in education — Actually, you're not often at dinner parties, frankly.(Laughter)If you work in education, you're not asked.(Laughter)And you're never asked back, curiously. That's strange to me. But if you are, and you say to somebody, you know, they say, "What do you do?" and you say you work in education, you can see the blood run from their face. They're like, "Oh my God," you know, "Why me?"(Laughter)"My one night out all week."(Laughter)But if you ask about their education, they pin you to the wall. Because it's one of those things that goes deep with people, am I right? Like religion, and money and other things. So I have a big interest in education, and I think we all do. We have a huge vested interest in it, partly because it's education that's meant to take us into this future that we can't grasp. If you think of it, children starting school this year will be retiring in 2065. Nobody has a clue, despite all the expertise that's been on parade for the past four days, what the world will look like in five years' time. And yet we're meant to be educating them for it. So the unpredictability, I think, is........	https://www.ted.com/talks/ken_robinson_says_schools_kill_creativity

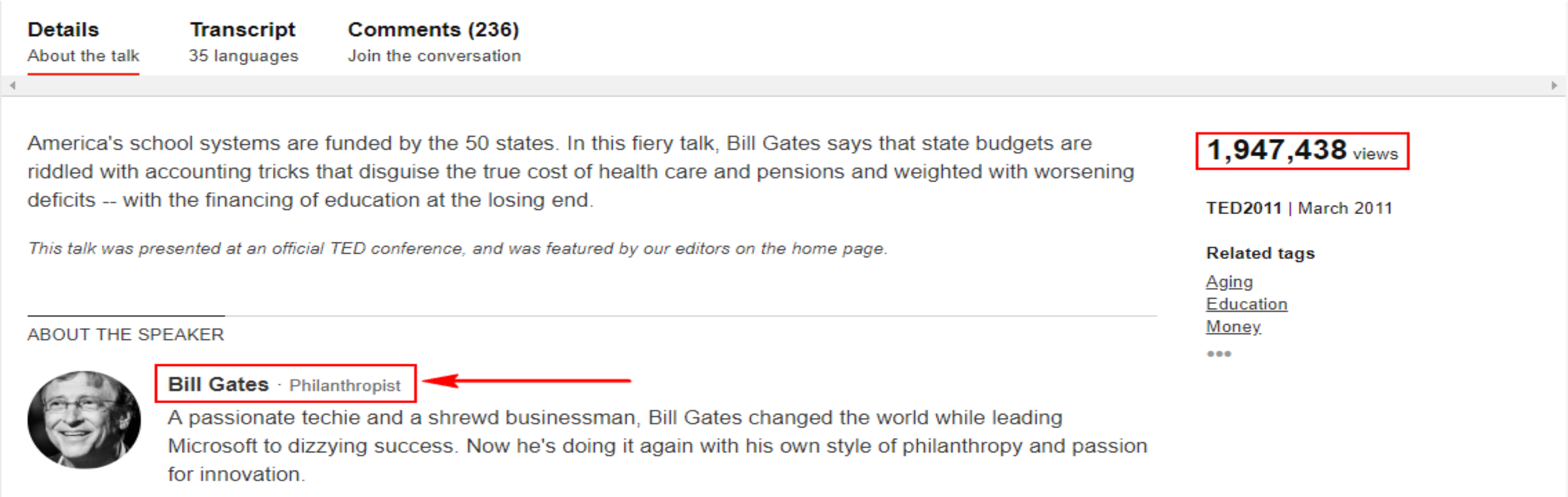
Data Understanding

Any Missing Values?

```
tedMain.speaker_occupation.fillna('UNKOWN', inplace=True)
```



Any outlier?



America's school systems are funded by the 50 states. In this fiery talk, Bill Gates says that state budgets are riddled with accounting tricks that disguise the true cost of health care and pensions and weighted with worsening deficits -- with the financing of education at the losing end.

This talk was presented at an official TED conference, and was featured by our editors on the home page.

ABOUT THE SPEAKER

Bill Gates · Philanthropist

A passionate techie and a shrewd businessman, Bill Gates changed the world while leading Microsoft to dizzying success. Now he's doing it again with his own style of philanthropy and passion for innovation.

1,947,438 views

TED2011 | March 2011

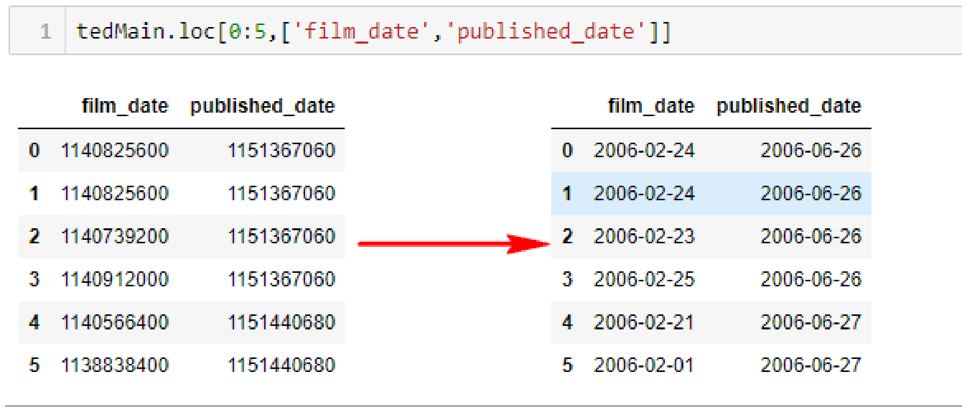
Related tags

Aging
Education
Money
•••

Views in million

Data Pre-Processing

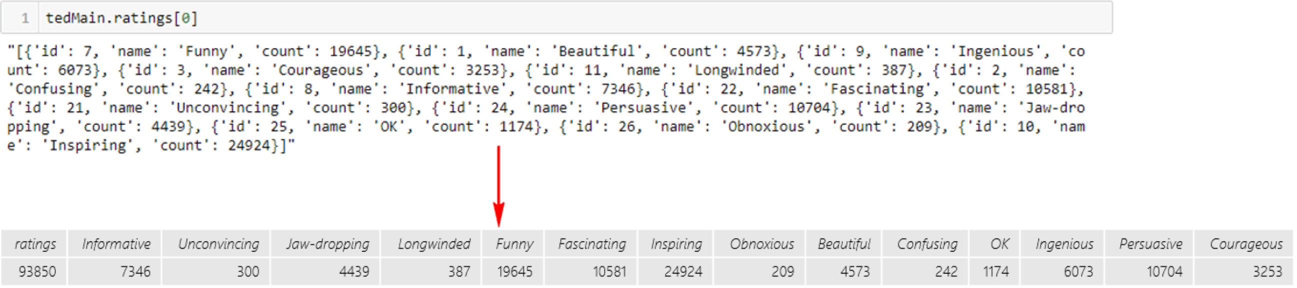
Data Normalization



```
tedMain.loc[0:5,['film_date','published_date']]
```

	film_date	published_date
0	1140825600	1151367060
1	1140825600	1151367060
2	1140739200	1151367060
3	1140912000	1151367060
4	1140566400	1151440680
5	1138838400	1151440680

	film_date	published_date
0	2006-02-24	2006-06-26
1	2006-02-24	2006-06-26
2	2006-02-23	2006-06-26
3	2006-02-25	2006-06-26
4	2006-02-21	2006-06-27
5	2006-02-01	2006-06-27

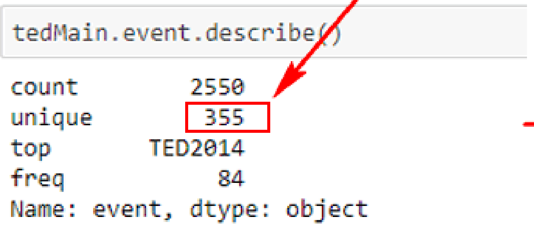


```
tedMain.ratings[0]
```

```
"[{'id': 7, 'name': 'Funny', 'count': 19645}, {'id': 1, 'name': 'Beautiful', 'count': 4573}, {'id': 9, 'name': 'Ingenious', 'co
unt': 6073}, {'id': 3, 'name': 'Courageous', 'count': 3253}, {'id': 11, 'name': 'Longwinded', 'count': 387}, {'id': 2, 'name':
'Confusing', 'count': 242}, {'id': 8, 'name': 'Informative', 'count': 7346}, {'id': 22, 'name': 'Fascinating', 'count': 10581},
{'id': 21, 'name': 'Unconvincing', 'count': 300}, {'id': 24, 'name': 'Persuasive', 'count': 10704}, {'id': 23, 'name': 'Jaw-dro
pping', 'count': 4439}, {'id': 25, 'name': 'OK', 'count': 1174}, {'id': 26, 'name': 'Obnoxious', 'count': 209}, {'id': 10, 'nam
e': 'Inspiring', 'count': 24924}]"
```

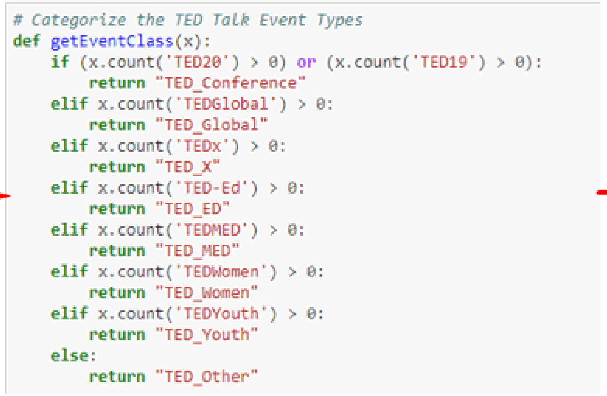
ratings	Informative	Unconvincing	Jaw-dropping	Longwinded	Funny	Fascinating	Inspiring	Obnoxious	Beautiful	Confusing	OK	Ingenious	Persuasive	Courageous
93850	7346	300	4439	387	19645	10581	24924	209	4573	242	1174	6073	10704	3253

Before



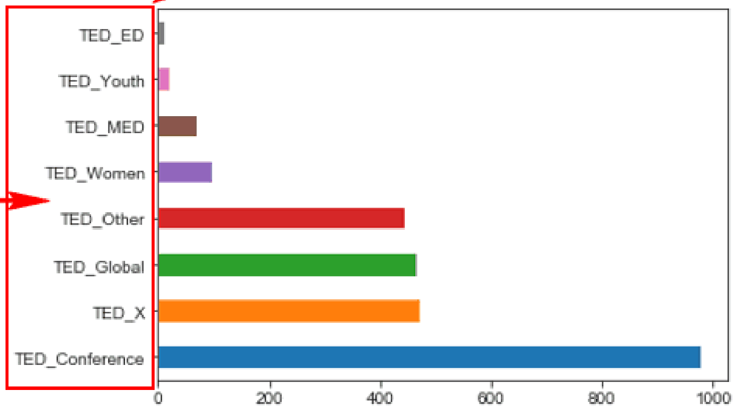
```
tedMain.event.describe()
```

```
count        2550
unique        355
top       TED2014
freq           84
Name: event, dtype: object
```



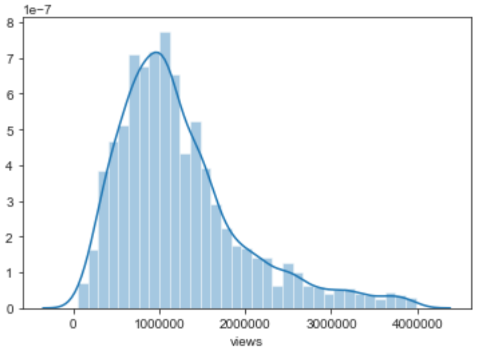
```
# Categorize the TED Talk Event Types
def getEventClass(x):
    if (x.count('TED20') > 0) or (x.count('TED19') > 0):
        return "TED_Conference"
    elif x.count('TEDGlobal') > 0:
        return "TED_Global"
    elif x.count('TEDx') > 0:
        return "TED_X"
    elif x.count('TED-Ed') > 0:
        return "TED_ED"
    elif x.count('TEDMED') > 0:
        return "TED_MED"
    elif x.count('TEDWomen') > 0:
        return "TED_Women"
    elif x.count('TEDYouth') > 0:
        return "TED_Youth"
    else:
        return "TED_Other"
```

After

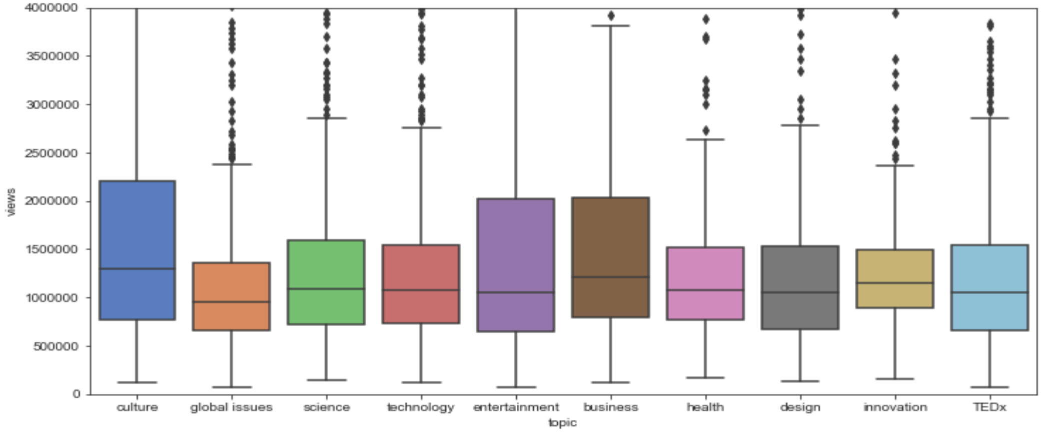


Data Pre-Processing

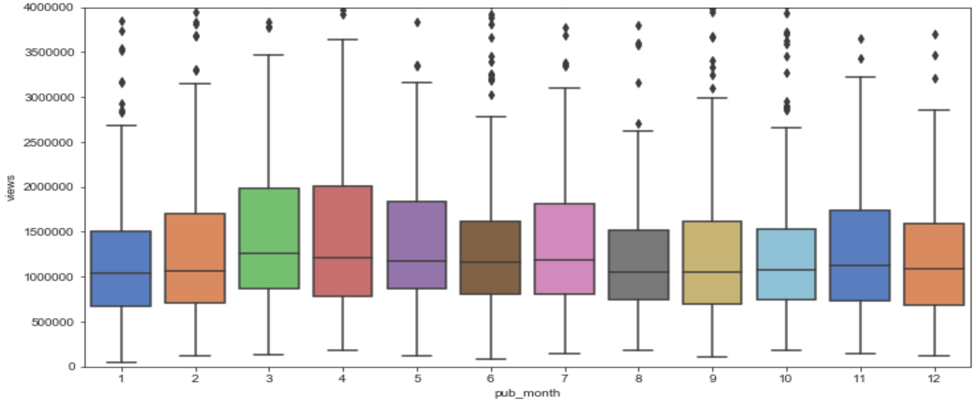
Data Visualization



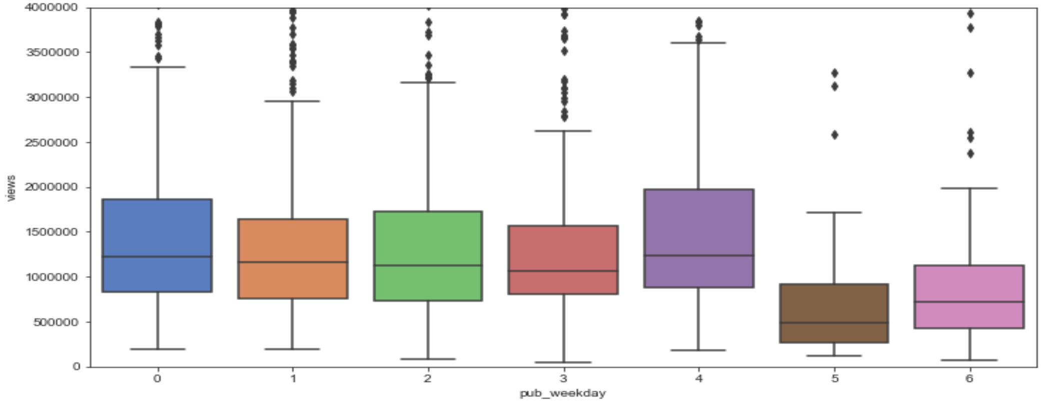
Distribution of views



Views Per Topic



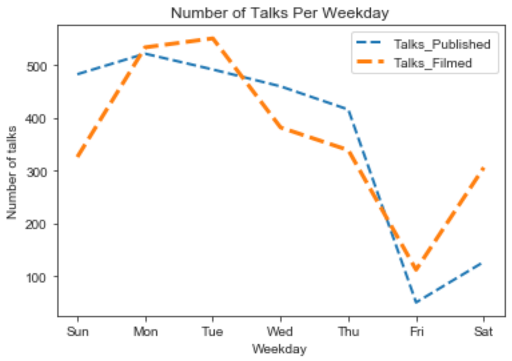
Views Per Publish Month



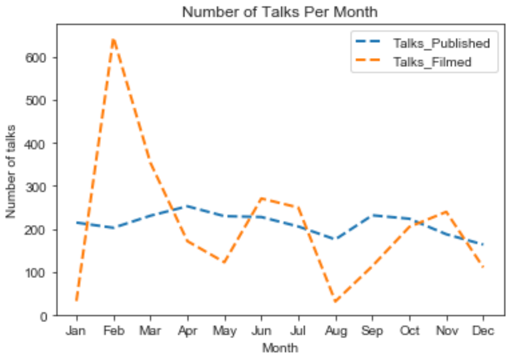
Views Per Publish Weekday

Data Pre-Processing

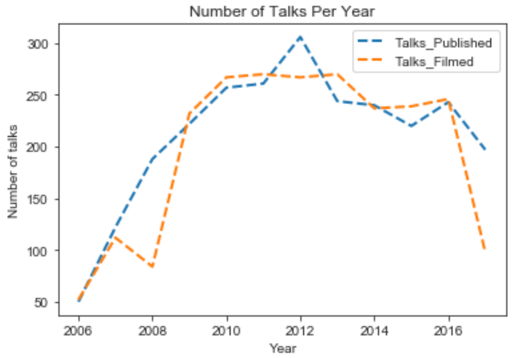
Data Visualization



Number of talks per weekday



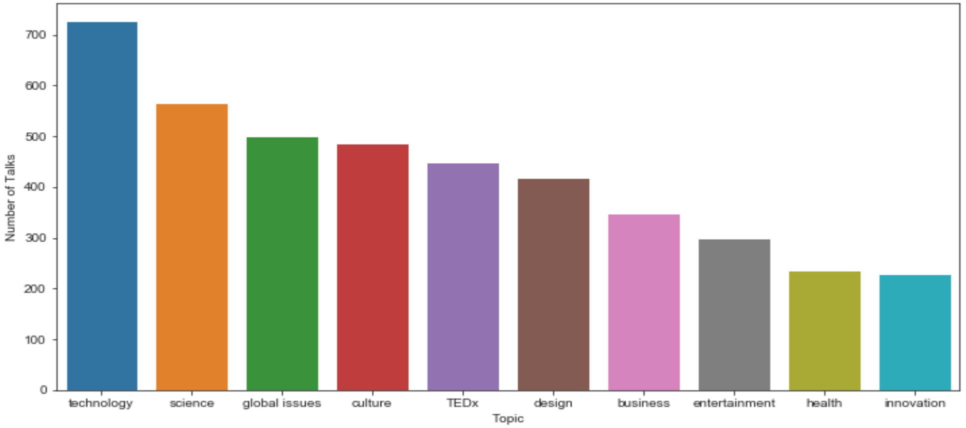
Number of talks per month



Number of talks per year

global issues'
culture'
technology'
TEDx'
design'
social change'
business'
science'
personal growth'
education'
storytelling'
TED Fellows'
communication'
entertainment'
science' technology'
community'
mental health'
language'
activism'
evolution'
innovation'
religion'
humor'
politics'
invention'
policy'
history'
writing'
comedy'
happiness'
Internet'
humanity'
identity'
visualizations'
psychology'
women'
children'
creativity'
health care'
health'
social media'
collaboration'
inequality'
life'
work'
future'
aging'
computers'
environment'
cities'
medical research'
climate change'
art'
brain'
philosophy'
potential'
Brain Trust'
cognitive science'
photography'
society'
change'
TED Brain
nature'
big problems

Word Cloud based on tags

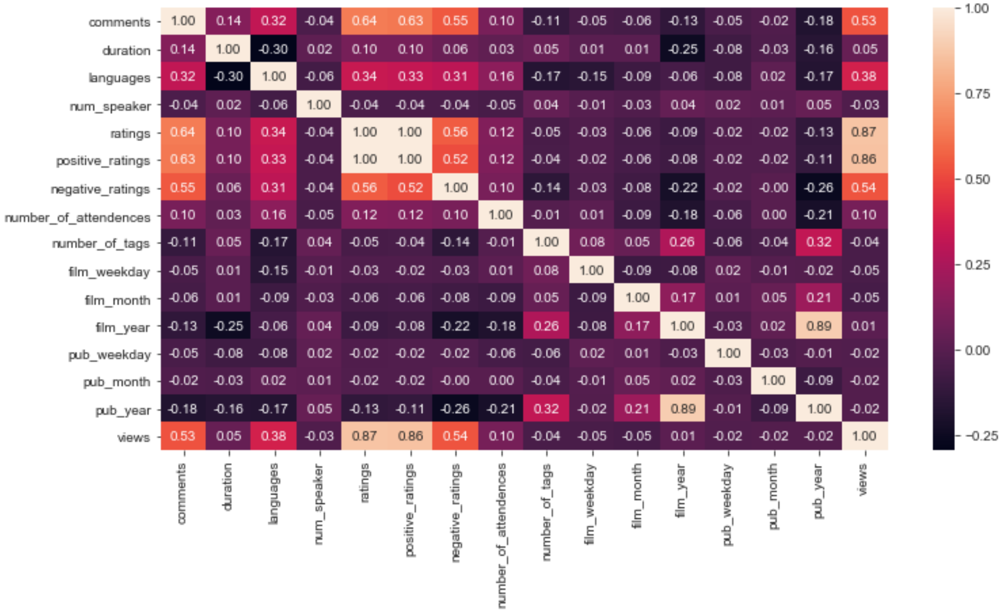


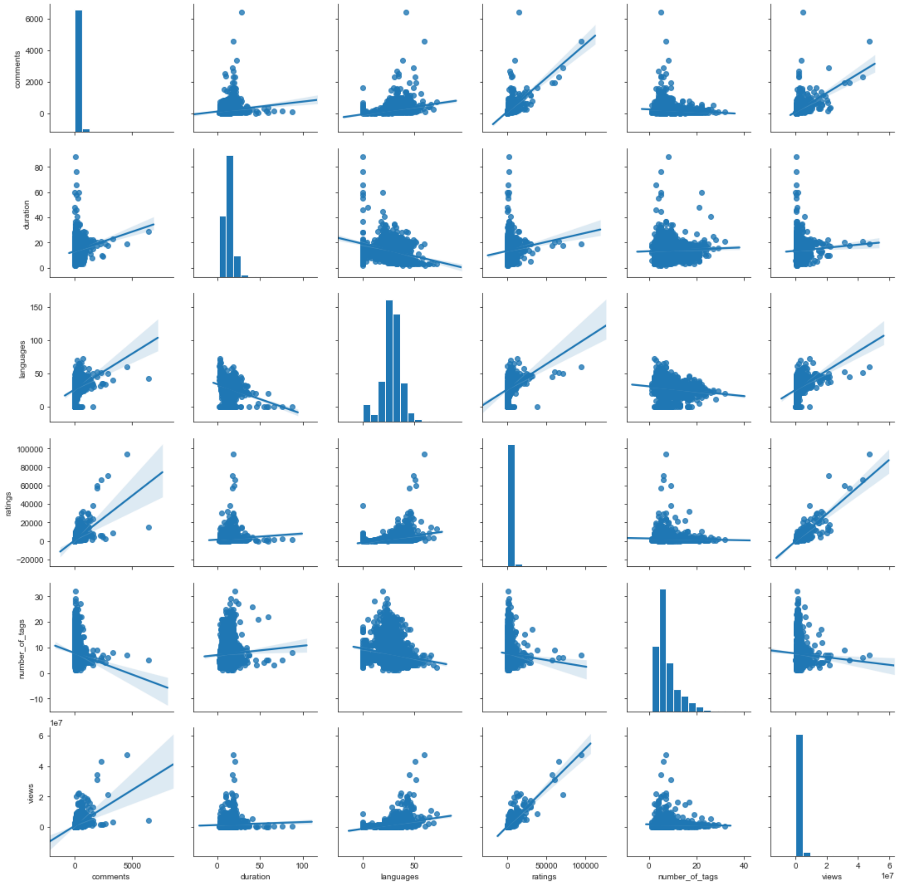
Number of talks per tag

Data Pre-Processing

Data Correlation

Heatmap with selected features based on EDA





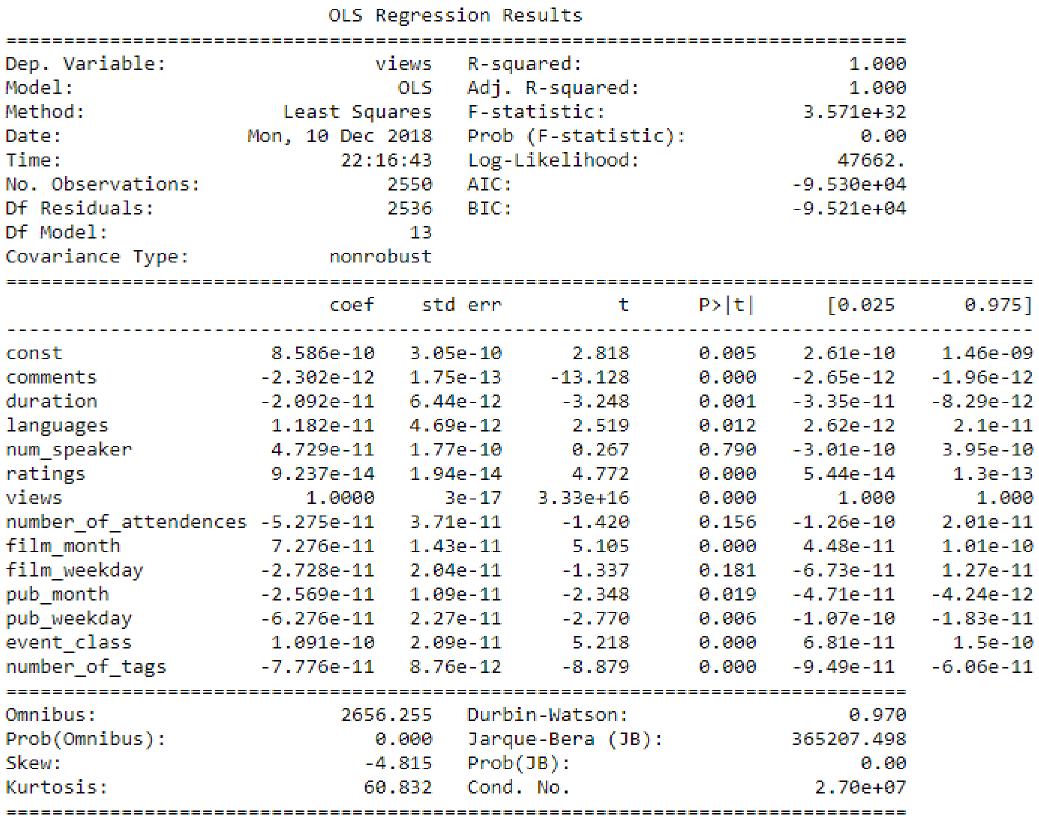
Regression visualization with selected features based on EDA

Model Building

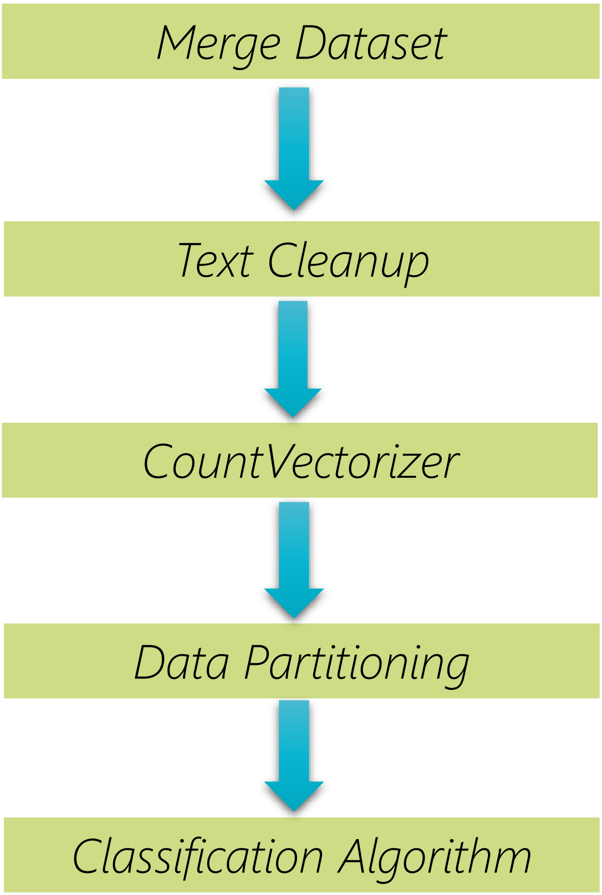
Classification Model – 'TED Main'

Classification Model – 'TED Transcript'

Linear Regression Analysis



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  views   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 3.571e+32
Date:                Mon, 10 Dec 2018   Prob (F-statistic):               0.00
Time:                        22:16:43   Log-Likelihood:                 47662.
No. Observations:                2550   AIC:                        -9.530e+04
Df Residuals:                    2536   BIC:                        -9.521e+04
Df Model:                          13
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                  8.586e-10   3.05e-10      2.818      0.005    2.61e-10    1.46e-09
comments              -2.302e-12   1.75e-13    -13.128      0.000   -2.65e-12   -1.96e-12
duration              -2.092e-11   6.44e-12     -3.248      0.001   -3.35e-11   -8.29e-12
languages              1.182e-11   4.69e-12      2.519      0.012    2.62e-12     2.1e-11
num_speaker            4.729e-11   1.77e-10      0.267      0.790   -3.01e-10    3.95e-10
ratings                9.237e-14   1.94e-14      4.772      0.000    5.44e-14     1.3e-13
views                     1.0000      3e-17   3.33e+16      0.000       1.000       1.000
number_of_attendences -5.275e-11   3.71e-11     -1.420      0.156   -1.26e-10    2.01e-11
film_month             7.276e-11   1.43e-11      5.105      0.000    4.48e-11    1.01e-10
film_weekday          -2.728e-11   2.04e-11     -1.337      0.181   -6.73e-11    1.27e-11
pub_month             -2.569e-11   1.09e-11     -2.348      0.019   -4.71e-11   -4.24e-12
pub_weekday           -6.276e-11   2.27e-11     -2.770      0.006   -1.07e-10   -1.83e-11
event_class            1.091e-10   2.09e-11      5.218      0.000    6.81e-11     1.5e-10
number_of_tags        -7.776e-11   8.76e-12     -8.879      0.000   -9.49e-11   -6.06e-11
==============================================================================
Omnibus:                     2656.255   Durbin-Watson:                   0.970
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           365207.498
Skew:                          -4.815   Prob(JB):                         0.00
Kurtosis:                      60.832   Cond. No.                     2.70e+07
==============================================================================
```



Merge Dataset

Text Cleanup

CountVectorizer

Data Partitioning

Classification Algorithm

Model Selection

Model Evaluation on Classification Model – 'TED Main'

Train/Test Spilt → 0.8 Train(2040) & 0.2 Test(510)

LogisticRegression

Classification Report	Precision	Recall	F1-Score	Support
False	0.78	0.80	0.79	279
True	0.75	0.74	0.74	231
Average/Total	0.77	0.77	0.77	510

Confusion Matrix	Predicted	
Actual	1	0
1	222	57
0	61	170

Apply RFE

DecisionTreeClassifier

Classification Report	Precision	Recall	F1-Score	Support
False	0.77	0.72	0.74	279
True	0.68	0.74	0.71	231
Average/Total	0.73	0.73	0.73	510

Confusion Matrix	Predicted	
Actual	1	0
1	200	79
0	61	170

RandomForestClassifier

Classification Report	Precision	Recall	F1-Score	Support
False	0.82	0.79	0.81	279
True	0.76	0.79	0.77	231
Average/Total	0.79	0.79	0.79	510

Confusion Matrix	Predicted	
Actual	1	0
1	227	58
0	49	182

Apply Scaling

Model Selection

Model Evaluation on Classification Model – 'TED Transcript'

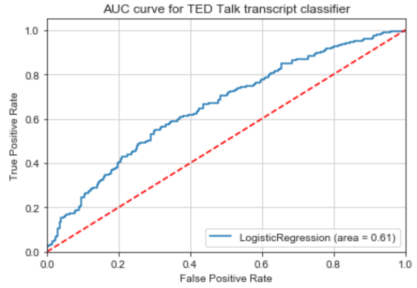
Train/Test Spilt

→ 0.8 Train(1973) & 0.2 Test(494)

LogisticRegression

Classification Report	Precision	Recall	F1-Score	Support
False	0.64	0.56	0.60	259
True	0.58	0.65	0.61	235
Average/Total	0.61	0.61	0.60	494

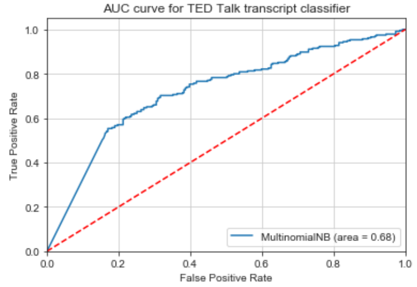
Confusion Matrix	Predicted	
Actual	1	0
1	146	113
0	82	153



MultinomialNB

Classification Report	Precision	Recall	F1-Score	Support
False	0.71	0.66	0.68	259
True	0.65	0.70	0.67	235
Average/Total	0.68	0.68	0.68	494

Confusion Matrix	Predicted	
Actual	1	0
1	170	89
0	70	165



Cross Validation

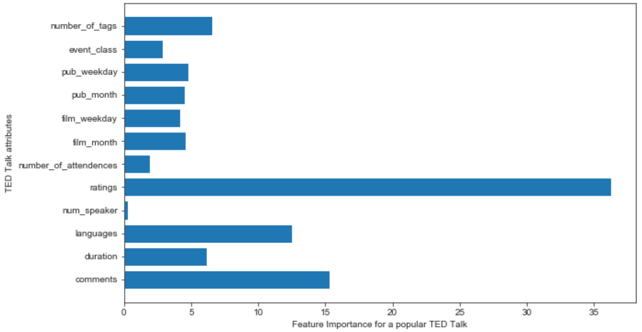
→ 10-Fold Cross Validation

	Logistic Regression	Multinomial NB
Mean Accuracy	60 %	65 %

Conclusion

1 - What makes a popular TED Talk?

Random Forest Classifier



Top 5 features selected by RFE (Recursive Feature Elimination)

- ***Comments***
- ***Languages***
- ***Ratings***
- ***Published Weekday***
- ***Number of tags***

2 - What are the most popular Topics that people like talk about?

- ***Technology***
- ***Culture***
- ***Global Issues***
- ***Design***
- ***Social change***

3 – Could the model predict the popularity of an TED Talk based on its transcript?

IT COULD! With a Prediction Accuracy of 68%.

Future Work

Apply Topic Modeling on text features

- In this project, there are some text features not being used while building the classification model, such as, **title**, **description**, and **tags** etc. due to time limitations. However, those features are actually more important.
- ***Proposed Algorithm:*** *Latent Dirichlet Allocation(LDA) is a popular algorithm for topic modeling with excellent implementations in the Python's Gensim package.*

Optimize the performance of Classification Models

- There are two kinds of classification models being built for this project, one is being trained based on the selected features from TED main dataset and the other one is trained based on the plain transcript of a talk. Neither model has HIGH prediction accuracy.
- Although the prediction accuracy depends on the dataset, planning further pre-pruning and post-post pruning to see if a better prediction accuracy can be obtained.

Thank You !

Q & A?

TED Ideas worth spreading