LET'S MAKE A HIGH-RATING MOVIE!

FEATURE ANALYSIS & RATING PREDICTION

Pengfei Hu, Xiaoxuan Lu, Siewying Gong, Kexian Chen, Junzhuo Gu


GUIDLINE

PROBLEM STATEMENT

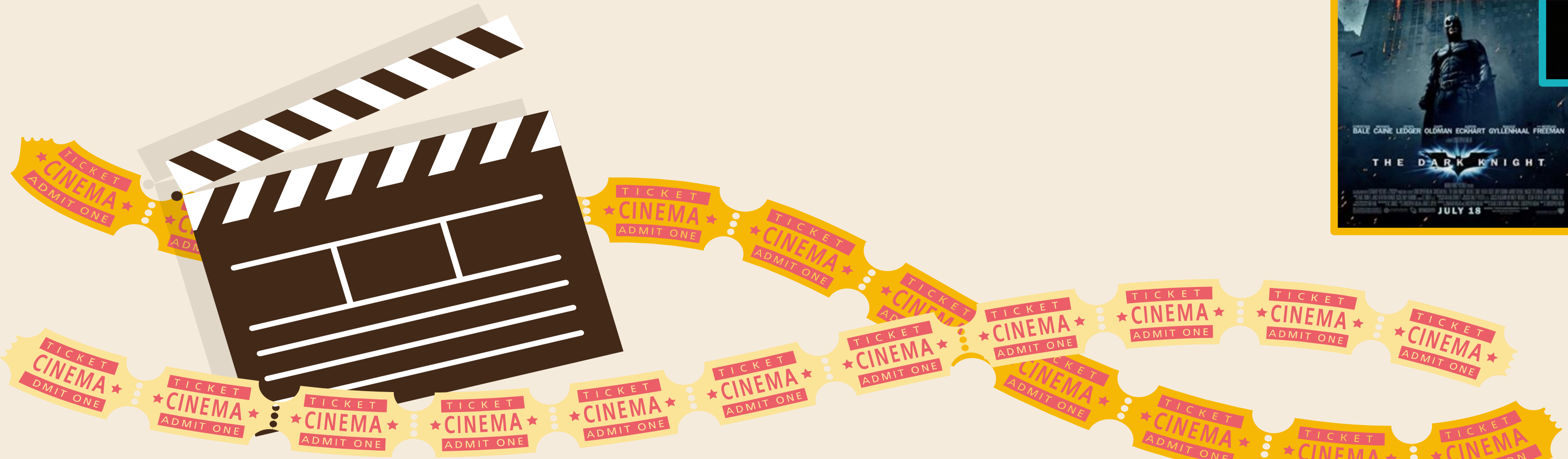
1. PROBLEM STATEMENT

Discover

a. What are important factors lead to **Movies' Rating**
b. Why is the rating like this? Is it because of the director? Release date? What is the underlying relationship?

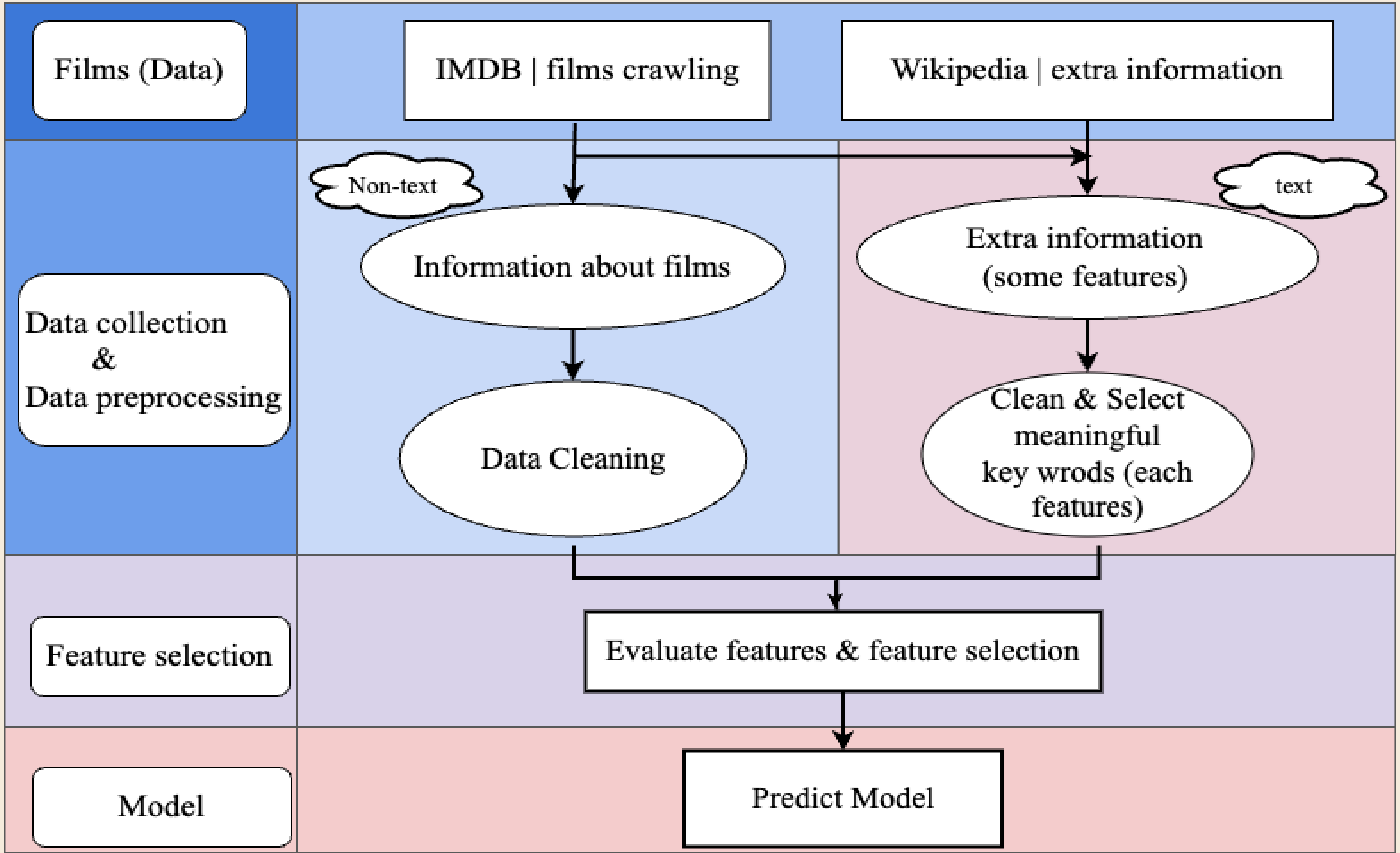
Rationalize

a. How do we know **what factors are Important**
b. Which model can help us predict the rating?

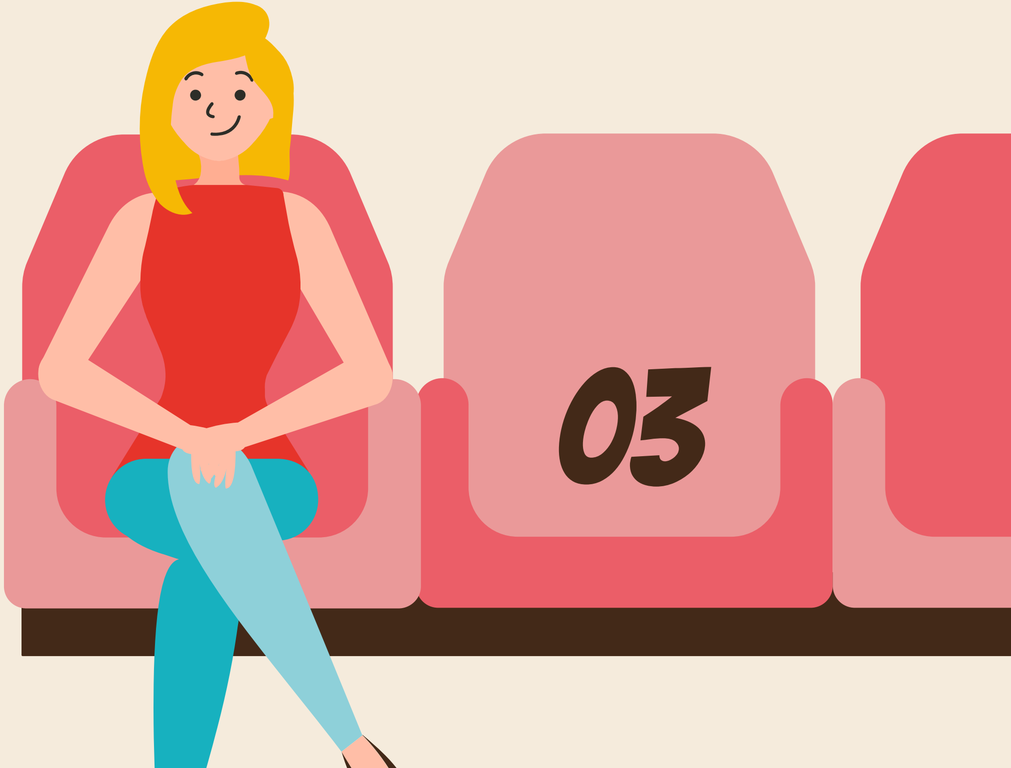


METHODOLOGY

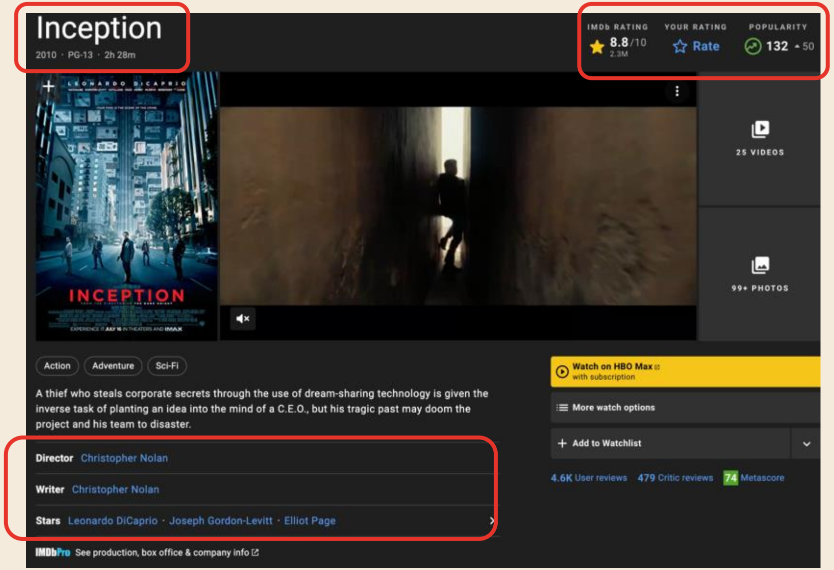
02. METHODOLOGY

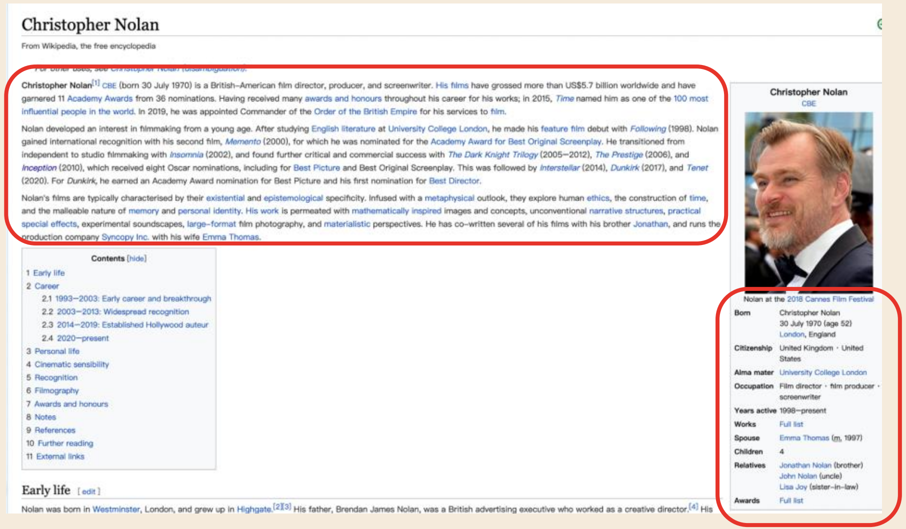


DATA COLLECTION



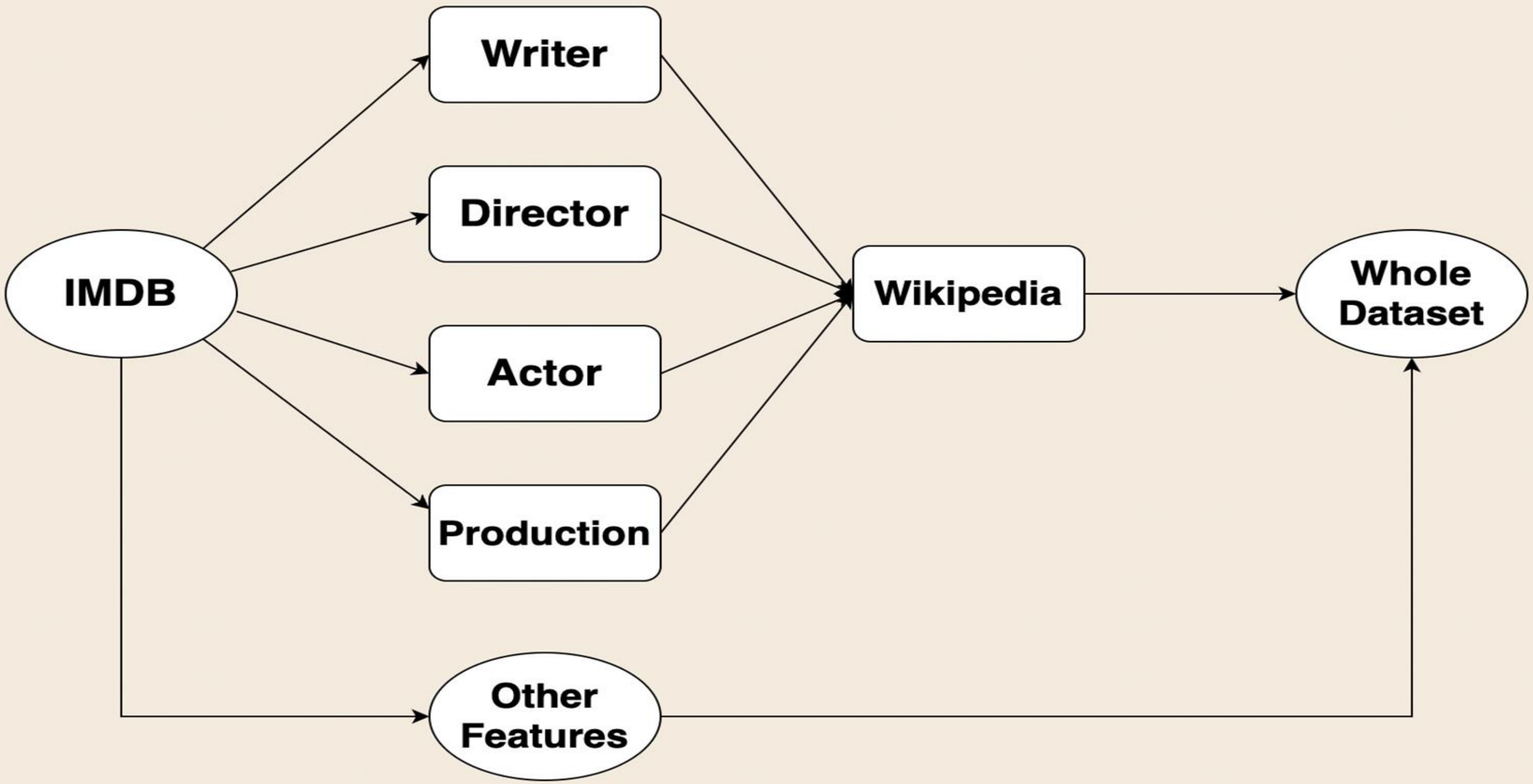
03. DATA COLLECTION - REAL TIME SCRAPING





- Basic Info Scrap from IMDB (***Self-Defined API*** using Beautiful Soup)
- Detail text scrap from Wikipedia (***Wiki API***)
- Size: **10158** rows , **30** columns
- Film year range: 1900 – 2022
- Number of vote **> 10K**

03. Data Collection - Structure



03. Data Collection – Initial Dataset (IMDB Part)

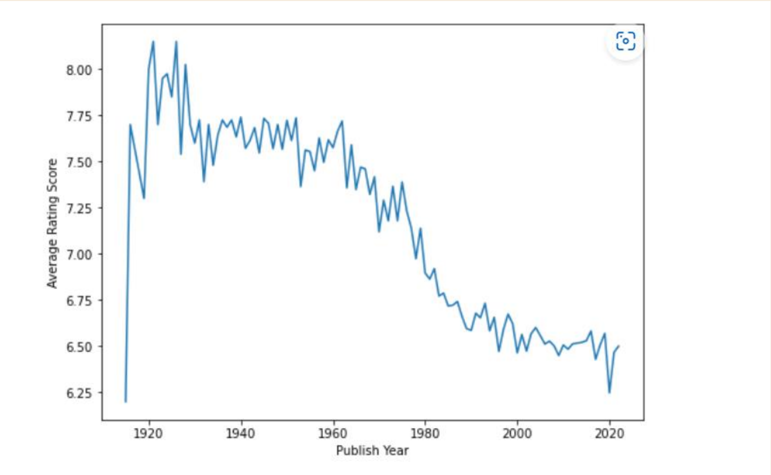
Data Information

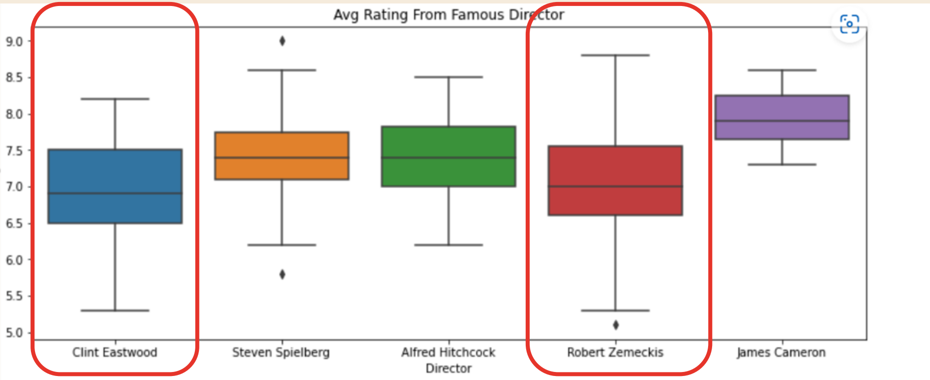
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10158 entries, 0 to 10157
Data columns (total 30 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   film_name             10158 non-null   object
 1   synopsis              10158 non-null   object
 2   genre_list            10158 non-null   object
 3   publish_year          10158 non-null   object
 4   MPAA                  9803 non-null    object
 5   Duration_minute       10158 non-null   object
 6   Rating                10158 non-null   object
 7   Rating_popularity     10158 non-null   object
 8   Popularity            4132 non-null    object
 9   Director              10158 non-null   object
 10  Writer                10158 non-null   object
 11  Stars                 10158 non-null   object
 12  Awards                8955 non-null    object
 13  User_reviews          10158 non-null   object
 14  Critic_reviews        10136 non-null   object
 15  Metascore             8140 non-null    object
 16  Release_date          10158 non-null   object
 17  Country_of_origin     10158 non-null   object
 18  Language              10158 non-null   object
 19  Filming_locations     9509 non-null    object
 20  Production_companies  10158 non-null   object
 21  Budget                0 non-null       object
 22  Gross_US_Canada       0 non-null       object
 23  Opening_weekend       0 non-null       object
 24  Gross_worldwide       0 non-null       object
 25  Runtime               10158 non-null   object
 26  Color                 9232 non-null    object
 27  Sound_mix             10158 non-null   object
 28  Aspect_ratio          9520 non-null    object
 29  film_url              10158 non-null   object
dtypes: object(30)
```

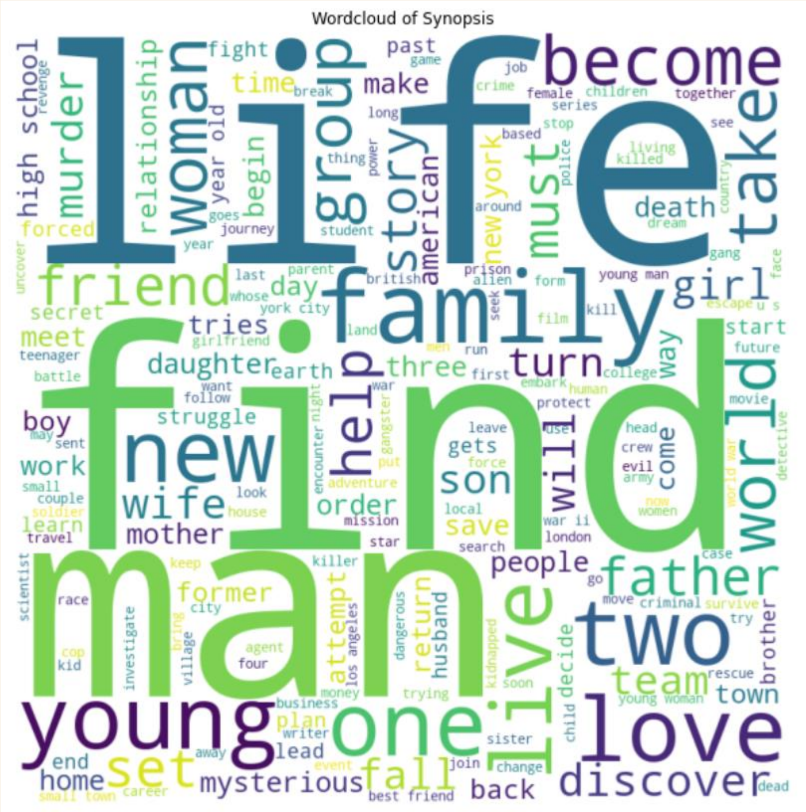
Data Overview

	film_name	synopsis	genre_list	publish_year	MPAA	Duration_minute	Rating	Rating_popularity	Popularity	Director	Writer	
0	The Birth of a Nation	The Stoneman family finds its friendship with ...	[Drama, History, War]	1915	TV-PG	195	6.2	25000	None	[D.W. Griffith]	[Thomas Dixon Jr., D.W. Griffith, Frank E. Woods]	[Lillian Mae M He Wa
1	Intolerance	The story of a poor young woman separated by p...	[Drama, History]	1916	Passed	197	7.7	16000	None	[D.W. Griffith]	[Hettie Grey Baker, Tod Browning, D.W. Griffith]	[Lillian F Harron M
2	Broken Blossoms	A frail waif, abused by her brutal boxer fathe...	[Drama, Romance]	1919	Not Rated	90	7.3	10000	None	[D.W. Griffith]	[Thomas Burke, D.W. Griffith]	[Lillian Ri Barthel D
3	The Cabinet of Dr. Caligari	Hypnotist Dr. Caligari uses a somnambulist, Ce...	[Horror, Mystery, Thriller]	1920	Not Rated	67	8.0	65000	2,409	[Robert Wiene]	[Carl Mayer, Hans Janowitz]	[W K C Fri
4	The Kid	The Tramp cares for an abandoned child, but ev...	[Comedy, Drama, Family]	1921	Passed	68	8.3	127000	None	[Charles Chaplin]	[Charles Chaplin]	[Cl Ch Purv Co
...	...	...	...	...	...	...	...	...	...	...	...	

DATA VISUALIZATION & INSIGHTS







03. DATA COLLECTION - INITIAL DATASET (WIKIPEDIA PART)

Text Overview

	actorext	directorext	writerext	productionext
0	\nLillian Diana Gish[1] (October 14, 1893 – Fe...	\nDavid Wark Griffith (January 22, 1875 – July...	Thomas Frederick Dixon Jr. (January 11, 1864 –...	The first of the AFI 100 Years... series of ci...
1	\nLillian Diana Gish[1] (October 14, 1893 – Fe...	\nDavid Wark Griffith (January 22, 1875 – July...	Tod Browning (born Charles Albert Browning Jr....	\nDavid Wark Griffith (January 22, 1875 – July...
2	\nLillian Diana Gish[1] (October 14, 1893 – Fe...	\nDavid Wark Griffith (January 22, 1875 – July...	\nThomas Burke (29 November 1886 – 22 Septembe...	\nDavid Wark Griffith (January 22, 1875 – July...
10153	\nSaravanan Sivakumar (born 23 July 1975), kno...	\nPandiraj (பாண்டிராஜ் in Tamil) is an Indian ...	\nA true toad is any member of the family Bufo...	\n\nSun Pictures is an Indian film distributio...
10154	\nKangana Ranaut is an Indian actress and film...		\nAnjolie Ela Menon (born 17 July 1940) is one...	\nThe Asylum is an American independent film c...
10155	\nEsha Deol (step-sister) \nVijay Singh Deol (...		\nJennifer Winget (born 30 May 1985) is an Ind...	\nDrishyam Films is an independent Indian film...
10156	\nCesar Manhilot[2] (born August 1, 1962), kno...	\nDarryl Yap (born January 7, 1987) is a Filip...	\nThis is a list of former and current politic...	\nViva Films is a Philippine film production c...
10157	\nLee Jung-jae (Korean: 이정재; born December 15,...	\nLee Jung-jae (Korean: 이정재; born December 15,...	\nBTS (Korean: 방탄소년단; RR: Bangtan Sonyeondan),...	

```
'\nLillian Diana Gish[1] (October 14, 1893\xa0– February 27, 1993) was an American actress,[2] director and screenwriter. Her film acting career spanned 75 years, from 1912, in silent film shorts, to 1987. Gish was called "The First Lady of American Cinema", and is credited with pioneering fundamental film performance techniques.[3] In 1999, the American Film Institute ranked Gish as the 17th greatest female movie star of classic Hollywood cinema.\nGish was a prominent film star from 1912 into the 1920s, being particularly associated with the films of director D. W. Griffith. This included her leading role in the highest-grossing film of the silent era, Griffith\'s The Birth of a Nation (1915). Her other major films and performances from the silent era are: Intolerance (1916), Broken Blossoms (1919), Way Down East (1920), Orphans of the Storm (1921), La Bohème (1926), and The Wind (1928).\nAt the dawn of the sound era, she returned to the stage and appeared in film infrequently, includi...'
```

04. DATA CLEANING AND FEATURE EVALUATION

NON-TEXT CLEANING

Format Normalization & Dummy

TEXT CLEANING

Meaningful Keywords modification & Dummy

EVALUATE FEATURES & FEATURE SELECTING

Statistical method

04. Data Cleaning - Non-Text Cleaning

- ***Format Normalization*** (remove $, convert to time to minutes, Datetime) & Transform to ***proper data type*** (①)
  - E.g object -> float
  - Modified some columns (ex. 'Date' Column → 'Year', 'Month', ***dummies***) (②)
- Drop features with many ***Nan values***
- Drop features which share the similar meaning with our labels and those unable to get before movie releasing (③)
  - Rating_popularity, User reviews

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10158 entries, 0 to 10157
Data columns (total 30 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   film_name        10158 non-null   object
 1   synopsis         10158 non-null   object
 2   genre_list       10158 non-null   object
 3   publish_year     10158 non-null   object
 4   MPAA             9803 non-null    object
 5   Duration_minute  10158 non-null   object
                        158 non-null   object
                        158 non-null   object
```

```
 4   MPAA                9803 non-null    object
 5   Duration_minute     10158 non-null   int64
 6   Rating              10158 non-null   object
 7   Rating_popularity   10158 non-null   int64
 8   Popularity          4132 non-null    object
 9   Director            10158 non-null   object
 10  Writer              10158 non-null   object
 11  Stars               10158 non-null   object
 12  User_reviews        10158 non-null   object
 13  Critic_reviews      10136 non-null   object
 14  Metascore           8140 non-null    object
 15  Release_date        10158 non-null   datetime64[ns]
 16  Country_of_origin   10158 non-null   object
 17  Language            10158 non-null   object
```

③ ①

```
data1['genre_list']

0             [Drama, History, War]
1                 [Drama, History]
2                 [Drama, Romance]
3      [Horror, Mystery, Thriller]
4          [Comedy, Drama, Family]
```

②

	gere_Action	gere_Adventure	gere_Animation	gere_Biography	gere_Comedy	gere_Crime	gere_Documentary	gere_Drama
0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	1
...	...	...	...	...	...	...	...	...
10153	1	0	0	0	0	1	0	1
10154	1	0	0	0	0	0	0	0

04. Data Cleaning - Text Cleaning (Labeling)

➔ Text cleaning:
  - Removing punctuation, numbers and stopwords
  - English words only

Before cleaning…

```
data['actorext'][0]
```

```
'\nLillian Diana Gish[1] (October 14, 1893\xa0– February 27, 1993) was an American actress,[2] director and screenwriter. Her film acting career spanned 75 years, from 1912, in silent film shorts, to 1987. Gish was called "The First Lady of American Cinema", and is credited with pioneering fundamental film performance techniques.[3] In 1999, the American Film Institute ranked Gish as the 17th greatest female movie star of classic Hollywood cinema.\nGish was a prominent film star from 1912 into the 1920s, being particularly associated with the films of director D. W. Griffith. This included her leading role in the highest-grossing film of the silent era, Griffith\'s The Birth of a Nation (1915). Her other major films and performances from the silent era are: Intolerance (1916), Broken Blossoms (1919), Way Down East (1920), Orphans of the Storm (1921), La Bohème (1926), and The Wind (1928).\nAt the dawn of the sound era, she returned to the stage and appeared in film infrequently, includi…'
```

After cleaning…

```
'Gish actress director screenwriter film acting career silent film shorts Gish First Lady Cinema fundamental film performance Film Institute ranked Gish female movie star classic cinema Gish prominent film star particularly associated director W included leading role highest film silent era Birth Nation major silent era Intolerance Broken Way East Storm La Wind dawn sound era returned stage film infrequently well known leading western Duel Sun thriller Night nominated Academy Award Best Supporting Actress former Gish major supporting Portrait Wedding Sweet Liberty considerable television work early closed career opposite film August later Gish advocate appreciation preservation silent film Despite better known film work accomplished stage Theater Hall Fame Academy Honorary Award career Center Honor contribution culture Mae Marsh born Mary Marsh film actress career Henry March June stage film actor Little Colonel W Birth Nation'
```

04. DATA CLEANING - TEXT CLEANING (KEY WORDS)

- ➔ Keywords extraction:
  - ◆ Semantic selection:
    - Autophrase （Open sourse）
    - Word frequency based

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv > cs > arXiv:1702.04457

Search... All fields Search
Help | Advanced Search

Computer Science > Computation and Language

[Submitted on 15 Feb 2017 (v1), last revised 11 Mar 2017 (this version, v2)]

Automated Phrase Mining from Massive Text Corpora

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, Jiawei Han

As one of the fundamental tasks in text analysis, phrase mining aims at extracting quality phrases from a text corpus. Phrase mining is important in various tasks such as information extraction/retrieval, taxonomy construction, and topic modeling. Most existing methods rely on complex, trained linguistic analyzers, and thus likely have unsatisfactory performance on text corpora of new domains and genres without extra but expensive adaption. Recently, a few data-driven methods have been developed successfully for extraction of phrases from massive domain-specific text. However, none of the state-of-the-art models is fully automated because they require human experts for designing rules or labeling phrases.
Since one can easily obtain many quality phrases from public knowledge bases to a scale that is much larger than that produced by human experts, in this paper, we

Download:
- PDF
- Other formats

(license)

Current browse context:
cs.CL
< prev | next >
new | recent | 1702
Change to browse by:
cs

References & Citations

Autophrase

```
dir_col
```

```
['film festival',
 'mystic river',
 'new york',
 'receive bafta',
 'best film',
 'beautiful mind',
 'black hawk',
 'television director',
 'horror genre',
 'special effects',
 'new wave',
 'film adaptation',
 'best picture',
 'director screenwriter',
 'award best',
```

04. Feature Selection & Evaluation

- ➢ Statistical Analysis for features:
  - ○ Select the features whose **p value is smaller than 0.05** in the **OLS model.**(Next slide)

OLS Regression Results

Dep. Variable:	Rating	**R-squared:**	0.437
Model:	OLS	**Adj. R-squared:**	0.401
Method:	Least Squares	**F-statistic:**	12.05
Date:	Mon, 21 Nov 2022	**Prob (F-statistic):**	0.00
Time:	01:11:05	**Log-Likelihood:**	-11100.
No. Observations:	10158	**AIC:**	2.343e+04
Df Residuals:	9542	**BIC:**	2.788e+04
Df Model:	615		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	36.1417	1.242	29.107	0.000	33.708	38.576
gere_Action	-0.2507	0.024	-10.550	0.000	-0.297	-0.204
gere_Adventure	-0.0449	0.027	-1.649	0.099	-0.098	0.008
gere_Animation	0.4628	0.046	10.016	0.000	0.372	0.553
gere_Biography	0.2993	0.034	8.892	0.000	0.233	0.365
gere_Comedy	-0.1367	0.023	-6.053	0.000	-0.181	-0.092
gere_Crime						045
gere_Documentary						22
gere_Drama						379
gere_Family						.100
gere_Fantasy						.058
gere_Film-Noir						229

	parameters	p_value
gere_Action	-2.507384e-01	7.0[illegible]7267e-26
gere_Animation	4.628226e-01	1.689157e-23
gere_Biography	2.992846e-01	7.078470e-19
gere_Horror	-4.886951e-01	1.507508e-59
gere_Music	1.546598e-02	7.395320e-01
oc_Canada	-1.091690e-01	3.703572e-04
lg_Spanish	1.938657e-02	4.601417e-01
Month_02	2.945559e+00	1.496685e-160
Month_03	3.001910e+00	8.537659e-168
Month_05	3.072038e+00	1.745274e-171
Month_06	3.032336e+00	3.363510e-172
Month_07	2.979984e+00	1.040504e-165
Month_08	2.963962e+00	7.309114e-165
Month_09	3.022873e+00	1.156099e-170
Month_10	3.030403e+00	4.722835e-172
Month_11	3.067446e+00	1.200848e-175
dir_receive bafta	1.044129e-15	8.915402e-02
dir_action film german film	1.686654e-16	7.874711e-01
act_superman	-3.790959e-16	9.684054e-02
act_black peral	-7.710922e-16	1.892491e-02

P_value<0.05

04. Data Cleaning - Feature Selection

- Genre
- MPAA
- Language
- Origin Country
- Release Location
- Release Date
- Sound Mix
- Director (text)
- Actor (text)
- Writer (text)
- Production company(text)
- Rating(Label)

04. Feature Selection & Evaluation

- Keep going:
  - All the selected features are useful for prediction model.
  - Text is useful to use but not optimal in this case, we could try different keywords extraction method next. (The features we have selected are regard as baseline)
  - A good prediction model will tell us the feature importance accordingly.

	parameters	p_value
gere_Action	−2.507384e−01	7.057267e−26
gere_Animation	4.628226e−01	1.689157e−23
gere_Biography	2.992846e−01	7.078470e−19
gere_Horror	−4.886951e−01	1.507508e−59
gere_Music	1.546598e−02	7.395320e−01
oc_Canada	−1.091690e−01	3.703572e−04
lg_Spanish	1.938657e−02	4.601417e−01
Month_02	2.945559e+00	1.496685e−160
Month_03	3.001910e+00	8.537659e−168
Month_05	3.072038e+00	1.745274e−171
Month_06	3.032336e+00	3.363510e−172
Month_07	2.979984e+00	1.040504e−165
Month_08	2.963962e+00	7.309114e−165
Month_09	3.022873e+00	1.156099e−170
Month_10	3.030403e+00	4.722835e−172
Month_11	3.067446e+00	1.200848e−175
dir_receive bafta	1.044129e−15	8.915402e−02
dir_action film german film	1.686654e−16	7.874711e−01
act_superman	−3.790959e−16	9.684054e−02
act_black peral	−7.710922e−16	1.892491e−02

MODEL

You could enter a subtitle here if you need it

05. MODEL - GUIDELINE ABOUT PREDICTION

Split Data

Training (80%)
5,689 rows

(5-Fold CV)
Validation (20%)
1,422 rows

70%

Test DataSet
3,047 rows

30%

Feature Selection

- Basis
- Basis + TF-IDF
- Basis + LDA
- Basis + Doc2Vec
- Basis + LDA + TF-IDF
- Basis + LDA + Doc2Vec
- Basis + TF-IDF + Doc2Vec
- Basis + TF-IDF + LDA + Doc2Vec

Regressor Options

- Linear Regression
- Gradient Boosting
- Ada Boost
- Random Forest
- Elastic Net
- Support Vector Regressor

Evaluation Metrics:

For Training:

- Adjusted R-Square

For Validation & Testing:

- MSE
- MAE

05. MODEL - LATENT FEATURES FROM TF-IDF

Synopsis (Storyline)

Actors' Description

Directors' Description

Writers' Description

Production Companies Description

Encoding to bag-of-words

Count Vectorizer
TF-IDF Transformer

N-gram within [1, 3]

Mainly Noun Entity

Max_feature = 2k

story_war	story_way	story_wife	story_woman	story_work	story_world
0.867676	0.0	0.0	0.000000	0.0	0.0
0.000000	0.0	0.0	0.478027	0.0	0.0
0.000000	0.0	0.0	0.000000	0.0	0.0
0.000000	0.0	0.0	0.000000	0.0	0.0
0.000000	0.0	0.0	0.000000	0.0	0.0
...	...	...	...	...	...
0.000000	0.0	0.0	0.614258	0.0	0.0
0.000000	0.0	0.0	0.000000	0.0	0.0
0.000000	0.0	0.0	0.000000	0.0	0.0
0.000000	0.0	0.0	0.000000	0.0	0.0
0.000000	0.0	0.0	0.000000	0.0	0.0

Encoding Demo Display

05. MODEL - LATENT FEATURES FROM LDA

Extra knowledge from Wiki

			writerext	productionext
			(January 11, 1864 –...	The first of the AFI 100 Years... series of ci...
			...bert Browning Jr....	\nDavid Wark Griffith (January 22, 1875 – July...
2	\nLillian Diana Gish[1] (October 14, 1893 – Fe...	\nDavid Wark Griffith (January 22, 1875 – July...	\nThomas Burke (29 November 1886 – 22 Septembe...	\nDavid Wark Griffith (January 22, 1875 – July...
3	\nWerner Johannes Krauss (Krauß in German; 23 ...	\nRobert Wiene (German: [ˈviːnə]; 27 April 187...	\nHans Janowitz (2 December 1890 – 25 May 1954...	Decla-Film (later Decla-Bioscop after 1920) wa...
4	Charles Spencer Chaplin III (May 5, 1925 – Mar...	\nSir Charles Spencer Chaplin Jr. KBE (16 Apri...	\nSir Charles Spencer Chaplin Jr. KBE (16 Apri...	\nSir Charles Spencer Chaplin Jr. KBE (16 Apri...
...	...	...	...	...
10153	\nSaravanan Sivakumar (born 23 July 1975), kno...	\nPandiraj (பாண்டிராஜ் in Tamil) is an Indian ...	\nA true toad is any member of the family Bufo...	\n\nSun Pictures is an Indian film distributio...
10154	\nKangana Ranaut is an Indian actress and film...		\nAnjolie Ela Menon (born 17 July 1940) is one...	\nThe Asylum is an American independent film c...
10155	\nEsha Deol (step-sister) \nVijay Singh Deol (...		\nJennifer Winget (born 30 May 1985) is an Ind...	\nDrishyam Films is an independent Indian film...
10156	\nCesar Manhilot[2] (born August 1, 1962), kno...	\nDarryl Yap (born January 7, 1987) is a Filip...	\nThis is a list of former and current politic...	\nViva Films is a Philippine film production c...
10157	\nLee Jung-jae (Korean: 이정재; born December 15,...	\nLee Jung-jae (Korean: 이정재; born December 15,...	\nBTS (Korean: 방탄소년단; RR: Bangtan Sonyeondan),...	

Fit Topics Model
(Latent Dirichlet Allocation)

Get topics and Related keywords
- ❑ Actor [8 Topics]
- ❑ Director [6 Topics]
- ❑ Writer [4 Topics]
- ❑ Production Company [6 Topics]

	actor_1	actor_2	actor_3	actor_4	actor_5	actor_6	actor_7	actor_8
0	0.292480	0.001945	0.001937	0.672852	0.001986	0.001941	0.001936	0.025238
1	0.287598	0.001892	0.001884	0.679199	0.001929	0.001885	0.001883	0.023361
2	0.270020	0.001678	0.001677	0.703125	0.001712	0.001681	0.001675	0.018433
3	0.690430	0.002613	0.002665	0.293457	0.002668	0.002609	0.002604	0.002647
4	0.003677	0.003645	0.003599	0.003653	0.836426	0.003597	0.003624	0.141602
...	...	...	...	...	...	...	...	...
10153	0.001267	0.001251	0.001251	0.010368	0.049072	0.001250	0.001249	0.934082
10154	0.000655	0.000648	0.000648	0.137085	0.000662	0.000648	0.000648	0.858887
						...33	0.001547	0.789551
						...16	0.001183	0.247559
						...53	0.001552	0.001594

Feature used in prediction

Topics Similarity

```
[(0,
  '0.024*"name" + 0.013*"surname" + 0.010*"output" + 0.008*"people" + '
  '0.008*"parser" + 0.007*"city" + 0.006*"century" + 0.006*"mw" + 0.005*"form" '
  '+ 0.005*"country" + 0.004*"family" + 0.004*"language" + 0.004*"word" + '
  '0.004*"origin" + 0.003*"world" + 0.003*"state" + 0.003*"part" + 0.003*"r" + '
  '0.003*"region" + 0.003*"l" + 0.002*"meaning" + 0.002*"population" + '
  '0.002*"citation" + 0.002*"saint" + 0.002*"son" + 0.002*"war" + 0.002*"e" + '
  '0.002*"church" + 0.002*"championship" + 0.002*"area"'),
 (1,
  '0.012*"state" + 0.008*"year" + 0.007*"time" + 0.007*"university" + '
  '0.006*"book" + 0.006*"series" + 0.006*"author" + 0.005*"york" + '
  '0.005*"member" + 0.005*"president" + 0.005*"school" + 0.005*"television" + '
  '0.005*"election" + 0.004*"season" + 0.004*"world" + 0.004*"show" + '
  '0.004*"john" + 0.004*"city" + 0.003*"party" + 0.003*"house" + '
  '0.003*"college" + 0.003*"medium" + 0.003*"life" + 0.003*"family" + '
  '0.003*"war" + 0.003*"company" + 0.003*"child" + 0.003*"law" + 0.002*"death" '
  '+ 0.002*"leader"'),
 (2,
                          "role" + '
                          ' + 0.008*"director" + '
                          04*"screenplay" + '
                          0.004*"debut" + '
                          04*"performance" + '
                          *"globe" + '
```

Topics Keywords

How to select a correct topics model?

Evaluation Metrics:
- Coherence Measures (C_v calculated by NPMI and cosine similarity)
- Our Domain Knowledge

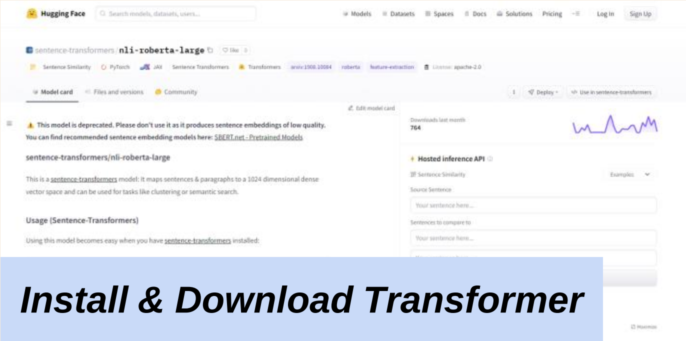
Parameter Tuning:
- Validation_set
- # of Topics
- Alpha & Beta value

	Validation_Set	Topics	Alpha	Beta	Coherence	var_name
334	75% Corpus	8	0.21	0.81	0.707183	Director
478	100% Corpus	10	0.01	0.61	0.702174	Director
369	75% Corpus	10	0.61	0.81	0.695168	Director
353	75% Corpus	10	0.01	0.61	0.692173	Director
279	75% Corpus	4	0.01	0.81	0.686236	Director
...	...	...	...	...	...	...
393	100% Corpus	2	0.61	0.61	0.534612	Director
389	100% Corpus	2	0.41	0.81	0.534232	Director
378	100% Corpus	2	0.01	0.61	0.529763	Director
					...3	Director
					...1	Director

Find optimal params

05. MODEL - LATENT FEATURES FROM DOC2VEC

BERT-Based Embedding Transformer



Install & Download Transformer

Fitting

Embedding Docs & Sentences

	actor_embed0	actor_embed1	actor_embed2	actor_embed3	actor_embed4	actor_embed5	actor_embed6	acto
0	0.776308	0.967113	-0.659960	0.308849	0.236675	0.239695	-0.720033	
1	0.776308	0.967113	-0.659960	0.308849	0.236675	0.239695	-0.720033	
2	0.776308	0.967113	-0.659960	0.308849	0.236675	0.239695	-0.720033	
3	-0.076159	1.587568	0.287104	0.479446	-0.732378	0.747140	-0.138448	
4	1.028658	0.558677	-0.531430	0.248476	-0.219392	0.523078	-0.892872	
...	...	...	...	...	...	...	...	
10153	1.150571	1.042381	-0.631035	0.122638	0.638389	0.087567	-1.478086	
10154	1.166747	0.972089	-0.850379	-0.077315	0.708110	0.211750	-0.183652	
10155	1.703650	0.839172	-0.597467	0.486304	0.196196	0.817926	-1.284960	
10156	1.294252	0.763948	0.146849	0.213496	-0.273649	0.425002	-0.649419	
10157	1.122336	0.964360	-0.216552	0.013807	-0.503666	0.621937	-1.210362	

10158 rows × 4096 columns

Embedding Result Display

Text Source (From Wikipedia)

- ❑ Actor description
- ❑ Director description
- ❑ Writer description
- ❑ Production Company description

[10,158 Docs × 4 different texts]

# of row equals to # of Docs;

# of columns equals to

- **4** entity (director, actor, production company, writer)
- **1,024** dimensions / per entity;

05. MODEL - FURTHER CONSIDERATION

OLS Regression Results

Dep. Variable:	Rating	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	3.109e+27
Date:	Sun, 20 Nov 2022	Prob (F-statistic):	0.00
Time:	06:50:51	Log-Likelihood:	2.9359e+05
No. Observations:	10158	AIC:	-5.859e+05
Df Residuals:	9541	BIC:	-5.815e+05
Df Model:	616		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.817e-14	1.22e-13	-0.149	0.882	-2.57e-13	2.21e-13
gere_Action	1.579e-14	2.25e-15	7.022	0.000	1.14e-14	2.02e-14
gere_Adventure	1.34e-16	2.56e-15	0.052	0.958	-4.89e-15	5.15e-15
gere_Animation	2.125e-14	4.37e-15	4.863	0.000	1.27e-14	2.98e-14
gere_Biography	-2.605e-14	3.18e-15	-8.193	0.000	-3.23e-14	-1.98e-14
gere_Comedy	2.738e-15	2.13e-15	1.286	0.198	-1.44e-15	6.91e-15
gere_Crime	2.815e-15	2.11e-15	1.335	0.182	-1.32e-15	6.95e-15
gere_Documentary	3.044e-15	5.83e-15	0.522	0.601	-8.38e-15	1.45e-14
gere_Drama	1.46e-15	2.07e-15	0.705	0.481	-2.6e-15	5.52e-15
gere_Family	-1.947e-15	4.03e-15	-0.484	0.629	-9.84e-15	5.94e-15
gere_Fantasy	-3.607e-15	2.99e-15	-1.207	0.228	-9.47e-15	2.25e-15
gere_Film-Noir	5.573e-15	1.01e-14	0.549	0.583	-1.43e-14	2.55e-14

From ML perspective:

Whether OLS regression will still have a **good predicting performance** for new launching films?

According to Baseline Model [70% Training + 30% Testing]

Seems like OLS cannot predict rating correctly

Mean Square Error: 2.662106e+19
Mean Absolute Error: 3.604561e+08

Why?

Two Possible Reason:

- **Under-fitting** {X uncorrelated with y}
  - -> **Solution:** Add latent features like TF-IDF, LDA or Embedding
- **Nonlinear & Co-linear relationships**
  - -> **Solution:** Use Non-linear & Tree Modeling like SVR, Boost etc..

05. MODEL - VERIFICATION OF REASONS

	Regressor Name	Attributes Set	Mean Absolute Error	Mean Square Error
4	LinearRegression	Basis	3.604561e+08	2.662106e+19
10	LinearRegression	Basis + TF-IDF	6.636086e+08	8.676807e+19
16	LinearRegression	Basis + LDA	1.302735e+08	2.785558e+18
22	LinearRegression	Basis + LDA + TF-IDF	1.360046e+08	3.144474e+18
28	LinearRegression	Basis + Doc2Vec	1.249400e+00	2.552300e+00
34	LinearRegression	Basis + LDA + Doc2Vec	1.236400e+00	2.503300e+00
40	LinearRegression	Basis + TF-IDF + Doc2Vec	1.186300e+00	2.298900e+00
46	LinearRegression	Basis + LDA + TF-IDF + Doc2Vec	1.175600e+00	2.252500e+00

When there is no embedding feature:
All MAE > 1 × e+8;
All MSE > 1 × e+19;

When it includes embedding features,
All MAE ∈ [1, 2];
All MSE ∈ [2, 3];

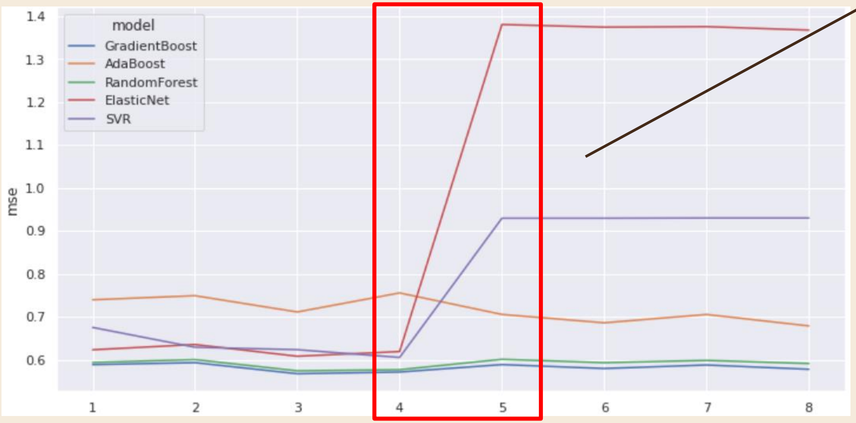
	Regressor Name	Attributes Set	Mean Absolute Error	Mean Square Error
0	GradientBoost	Basis	0.575371	0.588458
1	AdaBoost	Basis	0.676715	0.739542
2	RandomForest	Basis	0.577747	0.593447
3	ElasticNet	Basis	0.597413	0.623022
4	SVR	Basis	0.622784	0.675110

MSEs of alternative non-linear models fitting for basis attribute set are all within [0.5, 0.8], which is much smaller than the MSE (1×e+19) of the Linear Regression model.

Conclusion: Add more latent features from text and using non-linear regressor can actually help us get more precise and robust prediction

05. MODEL - EVALUATION

SVR and **Elastic Net** have instant changes based on the Doc2Vec features



Attributes Set	GradientBoost	AdaBoost	RandomForest	ElasticNet	SVR
Basis	0.58846	0.73954	0.59345	0.62302	0.67511
Basis + TF-IDF	0.59303	0.74908	0.60013	0.63529	0.62891
Basis + LDA	0.56721	0.71113	0.57398	0.60798	0.62323
Basis + LDA + TF-IDF	0.57120	0.75551	0.57661	0.61890	0.60518
Basis + WIKI	0.58854	0.70541	0.60081	1.38101	0.92961
Basis + LDA + WIKI	0.57939	0.68590	0.59297	1.37487	0.92963
Basis + TF-IDF + WIKI	0.58758	0.70532	0.59832	1.37606	0.93008
Basis + LDA + TF-IDF + WIKI	0.57762	0.67865	0.59101	1.36799	0.93010

- **5** candidates **non-linear** classifiers
- **8** different selections among latent factors based on **wiki text**:
  - **Topics Similarity** - LDA
  - **TF-IDF** - max_features = 2k
  - **Doc2Vec** - BERT Transformer
- Focus more on **MSE** for evaluation and model comparison
- Computationally intensive

Gradient Boosting has relatively lower *MSE* compared with **Elastic Net**

05. MODEL - FIND TOP MODEL FOR PREDICTION

	Regressor Name	Attributes Set	Mean Absolute Error	Mean Square Error	parameter
0	GradientBoost	Basis + LDA	0.5688	0.5672	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
1	GradientBoost	Basis + LDA + TF-IDF	0.5705	0.5712	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
2	RandomForest	Basis + LDA	0.5721	0.5740	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
3	RandomForest	Basis + LDA + TF-IDF	0.5715	0.5766	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
4	GradientBoost	Basis + LDA + TF-IDF + Doc2Vec	0.5751	0.5776	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
5	GradientBoost	Basis + LDA + Doc2Vec	0.5756	0.5794	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
6	GradientBoost	Basis + TF-IDF + Doc2Vec	0.5802	0.5876	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
7	GradientBoost	Basis	0.5754	0.5885	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
8	GradientBoost	Basis + Doc2Vec	0.5806	0.5885	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
9	RandomForest	Basis + LDA + TF-IDF + Doc2Vec	0.5812	0.5910	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
10	GradientBoost	Basis + TF-IDF	0.5776	0.5930	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...
11	RandomForest	Basis + LDA + Doc2Vec	0.5816	0.5930	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
12	RandomForest	Basis	0.5777	0.5934	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
13	RandomForest	Basis + TF-IDF + Doc2Vec	0.5846	0.5983	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
14	RandomForest	Basis + TF-IDF	0.5775	0.6001	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
15	RandomForest	Basis + Doc2Vec	0.5861	0.6008	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...
16	SVR	Basis + LDA + TF-IDF	0.5903	0.6052	{'C': 100, 'cache_size': 200, 'coef0': 0.0, 'd...
17	ElasticNet	Basis + LDA	0.5907	0.6080	{'alpha': 1.0, 'copy_X': True, 'fit_intercept'...
18	ElasticNet	Basis + LDA + TF-IDF	0.5980	0.6189	{'alpha': 1.0, 'copy_X': True, 'fit_intercept'...
19	ElasticNet	Basis	0.5974	0.6230	{'alpha': 1.0, 'copy_X': True, 'fit_intercept'...
20	SVR	Basis + LDA	0.5988	0.6232	{'C': 100, 'cache_size': 200, 'coef0': 0.0, 'd...
21	SVR	Basis + TF-IDF	0.6012	0.6289	{'C': 100, 'cache_size': 200, 'coef0': 0.0, 'd...
22	ElasticNet	Basis + TF-IDF	0.6051	0.6353	{'alpha': 1.0, 'copy_X': True, 'fit_intercept'...
23	SVR	Basis	0.6228	0.6751	{'C': 100, 'cache_size': 200, 'coef0': 0.0, 'd...
24	AdaBoost	Basis + LDA + TF-IDF + Doc2Vec	0.6419	0.6787	{'base_estimator': None, 'learning_rate': 1.0,...
25	AdaBoost	Basis + LDA + Doc2Vec	0.6446	0.6859	{'base_estimator': None, 'learning_rate': 1.0,...
26	AdaBoost	Basis + TF-IDF + Doc2Vec	0.6526	0.7053	{'base_estimator': None, 'learning_rate': 1.0,...
27	AdaBoost	Basis + Doc2Vec	0.6551	0.7054	{'base_estimator': None, 'learning_rate': 1.0,...
28	AdaBoost	Basis + LDA	0.6647	0.7111	{'base_estimator': None, 'learning_rate': 1.0,...
29	AdaBoost	Basis	0.6767	0.7395	{'base_estimator': None, 'learning_rate': 1.0,...
30	AdaBoost	Basis + TF-IDF	0.6839	0.7491	{'base_estimator': None, 'learning_rate': 1.0,...

Here is the **ascending rank** according to **MSE** for each regressor and each attribute set

- **Top 2 optimal regressor**
  - Gradient Boosting Regressor
  - Random Forest Regressor
- **Top 2 optimal attributes set**
  - Basis + LDA
  - Basis + LDA + TF-IDF

Then we can use these to find the best model with optimal parameters

05. MODEL – 5-FOLD CV FOR GRADIENT BOOST:

Top 10 Gradient Boost by [Basis + LDA] Features

	params	mean_valid_score	std_valid_score	rank_valid_score	split0_valid_score	split1_valid_score	split2_valid_score	split3_valid_score	split4_valid_score
0	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546286	0.041027	1	-0.546919	-0.607992	-0.571250	-0.493306	-0.511966
1	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546286	0.041027	1	-0.546919	-0.607992	-0.571250	-0.493306	-0.511966
2	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546356	0.040731	3	-0.544339	-0.607886	-0.572684	-0.495448	-0.511426
3	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546356	0.040731	3	-0.544339	-0.607886	-0.572684	-0.495448	-0.511426
4	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546431	0.044475	5	-0.546176	-0.608657	-0.581314	-0.490554	-0.505455
5	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546434	0.044463	6	-0.546135	-0.608652	-0.581314	-0.490554	-0.505515
6	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546508	0.042594	7	-0.544712	-0.607888	-0.576936	-0.488780	-0.514225
7	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546516	0.042582	8	-0.544681	-0.607883	-0.576936	-0.488780	-0.514299
8	{'criterion': 'friedman_mse', 'learning_rate':...	-0.547387	0.042075	9	-0.542515	-0.610633	-0.575998	-0.493267	-0.514521
9	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.547387	0.042075	9	-0.542515	-0.610633	-0.575998	-0.493267	-0.514521

Top 10 Gradient Boost by [Basis + LDA + TF-IDF] Features

	params	mean_valid_score	std_valid_score	rank_valid_score	split0_valid_score	split1_valid_score	split2_valid_score	split3_valid_score	split4_valid_score
0	{'criterion': 'friedman_mse', 'learning_rate':...	-0.555572	0.040309	1	-0.551565	-0.611716	-0.581745	-0.491874	-0.540962
1	{'criterion': 'mse', 'learning_rate': 0.1, 'ma...	-0.555612	0.040289	2	-0.551600	-0.611842	-0.581745	-0.492090	-0.540783
2	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.556961	0.041422	3	-0.553420	-0.615110	-0.585482	-0.493990	-0.536803
3	{'criterion': 'friedman_mse', 'learning_rate':...	-0.557033	0.041398	4	-0.553840	-0.615048	-0.585482	-0.493990	-0.536803
4	{'criterion': 'friedman_mse', 'learning_rate':...	-0.557238	0.039516	5	-0.552031	-0.614502	-0.581576	-0.496785	-0.541295
5	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.557259	0.039521	6	-0.552197	-0.614502	-0.581532	-0.496723	-0.541341
6	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.557920	0.042659	7	-0.558944	-0.615469	-0.588583	-0.492513	-0.534094
7	{'criterion': 'friedman_mse', 'learning_rate':...	-0.557996	0.042648	8	-0.559370	-0.615420	-0.588583	-0.492513	-0.534094
8	{'criterion': 'friedman_mse', 'learning_rate':...	-0.558140	0.044888	9	-0.554243	-0.621008	-0.587723	-0.488259	-0.539467
9	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.558173	0.041364	10	-0.553555	-0.613986	-0.594335	-0.502342	-0.526650

- ❑ Tuning Model:
  - ❑ Gradient Boosting Regressor
- ❑ Attribute Sets:
  - ❑ Basis + LDA
  - ❑ Basis + LDA + TF-IDF
- ❑ Scoring Metrics:
  - ❑ **Negative MSE**
- ❑ Hyperparameters Used:

```
[9] gradientboost_parameters = {
      'learning_rate': [0.05, 0.1, 0.2, 0.5],
      'n_estimators': [50, 100, 200],
      'criterion': ['friedman_mse', 'mse'],
      'min_samples_split': [2, 5, 10],
      'max_depth': [3, 5]
    }
```

05. MODEL – 5-FOLD CV FOR RANDOM FOREST:

Top 10 Random Forest by [Basis + LDA] Features

	params	mean_valid_score	std_valid_score	rank_valid_score	split0_valid_score	split1_valid_score	split2_valid_score	split3_valid_score	split4_valid_score
0	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.576233	0.040229	1	-0.559025	-0.635505	-0.606823	-0.521098	-0.558713
1	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.576292	0.040258	2	-0.559526	-0.635209	-0.606969	-0.520410	-0.559345
2	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.576327	0.040062	3	-0.559851	-0.634450	-0.608335	-0.521639	-0.557357
3	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.576610	0.040209	4	-0.560380	-0.635735	-0.607245	-0.521369	-0.558321
4	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.576806	0.040211	5	-0.560131	-0.636044	-0.607450	-0.521781	-0.558621
5	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.576899	0.039522	6	-0.561012	-0.634671	-0.608343	-0.523762	-0.556707
6	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.580448	0.038366	7	-0.563673	-0.639835	-0.606977	-0.530751	-0.561003
7	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.580698	0.038843	8	-0.562092	-0.641750	-0.605382	-0.529700	-0.564564
8	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.581172	0.038658	9	-0.562126	-0.641705	-0.606054	-0.530326	-0.565651
9	{'max_depth': 5, 'min_samples_split': 2, 'n_es...	-0.598420	0.043189	10	-0.585885	-0.660630	-0.631210	-0.536893	-0.577482

Top 10 Random Forest by [Basis + LDA + TF-IDF] Features

	params	mean_valid_score	std_valid_score	rank_valid_score	split0_valid_score	split1_valid_score	split2_valid_score	split3_valid_score	split4_valid_score
0	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.579482	0.037778	1	-0.560561	-0.632049	-0.611619	-0.526366	-0.566818
1	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.579805	0.037833	2	-0.561652	-0.632859	-0.610909	-0.525986	-0.567618
2	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.579825	0.037861	3	-0.560465	-0.633177	-0.611296	-0.526932	-0.567256
3	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.580024	0.038295	4	-0.560366	-0.633637	-0.613206	-0.527523	-0.565388
4	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.580064	0.038189	5	-0.561965	-0.634301	-0.611347	-0.526713	-0.565995
5	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.580262	0.038583	6	-0.560605	-0.635316	-0.611990	-0.527125	-0.566271
6	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.581851	0.036374	7	-0.559771	-0.634477	-0.611744	-0.533766	-0.569495
7	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.581900	0.035806	8	-0.562078	-0.634788	-0.609522	-0.534011	-0.569103
8	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.582454	0.036942	9	-0.559382	-0.636648	-0.611346	-0.533491	-0.571405
9	{'max_depth': 5, 'min_samples_split': 5, 'n_es...	-0.600252	0.040404	10	-0.586491	-0.656486	-0.634071	-0.542650	-0.581562

- Tuning Model:
  - Random Forest Regressor
- Attribute Sets:
  - Basis + LDA
  - Basis + LDA + TF-IDF
- Scoring Metrics:
  - **Negative MSE**
- Hyperparameters Used:

```
[27] randomforest_parameters = {
       'n_estimators': [50, 100, 200],
       'criterion': ['squared_error', 'absolute_error'],
       'max_depth': [3, 5, 7],
       'max_features': ['sqrt', 'log2', 'auto'],
       'min_samples_split': [2, 5, 10],
     }
```

06. MODEL - THE BEST MODEL:

Top 10 Rank by Negative MSE among Gradient Boosting

	Regressor Name	Attribute Set	params	mean_valid_score	std_valid_score	rank_valid_score
0	GradientBoost	Basis+LDA	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546286	0.041027	1
1	GradientBoost	Basis+LDA	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546286	0.041027	1
2	GradientBoost	Basis+LDA	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546356	0.040731	3
3	GradientBoost	Basis+LDA	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546356	0.040731	3
4	GradientBoost	Basis+LDA	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546431	0.044475	5
5	GradientBoost	Basis+LDA	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546434	0.044463	6
6	GradientBoost	Basis+LDA	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.546508	0.042594	7
7	GradientBoost	Basis+LDA	{'criterion': 'friedman_mse', 'learning_rate':...	-0.546516	0.042582	8
8	GradientBoost	Basis+LDA	{'criterion': 'friedman_mse', 'learning_rate':...	-0.547387	0.042075	9
9	GradientBoost	Basis+LDA	{'criterion': 'mse', 'learning_rate': 0.05, 'm...	-0.547387	0.042075	9

Top 10 Rank by Negative MSE among Random Forest

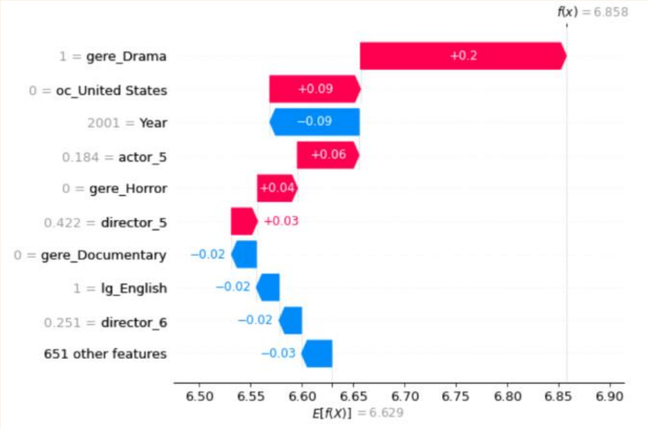
	Regressor Name	Attribute Set	params	mean_valid_score	std_valid_score	rank_valid_score
0	Random Forest	Basis+LDA	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.576233	0.040229	1
1	Random Forest	Basis+LDA	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.576292	0.040258	2
2	Random Forest	Basis+LDA	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.576327	0.040062	3
3	Random Forest	Basis+LDA	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.576610	0.040209	4
4	Random Forest	Basis+LDA	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.576806	0.040211	5
5	Random Forest	Basis+LDA	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.576899	0.039522	6
6	Random Forest	Basis+LDA+TF-IDF	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.579482	0.037778	1
7	Random Forest	Basis+LDA+TF-IDF	{'max_depth': 7, 'min_samples_split': 10, 'n_e...	-0.579805	0.037833	2
8	Random Forest	Basis+LDA+TF-IDF	{'max_depth': 7, 'min_samples_split': 2, 'n_es...	-0.579825	0.037861	3
9	Random Forest	Basis+LDA+TF-IDF	{'max_depth': 7, 'min_samples_split': 5, 'n_es...	-0.580024	0.038295	4

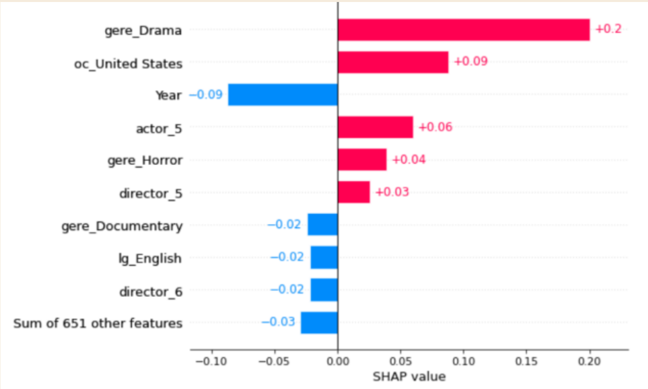
According to CV trails:

- The prediction of Gradient boosting regressor is significantly **better** than results of Random Forest Regressor
- The best model will be Gradient boosting regressor fitting with [Basis + LDA] attributes set and related optimal parameter is:

{'max_depth': 7, 'min_samples_split': 5, 'n_estimators': 200, 'criterion': 'squared_error', 'max_feature': 'auto'}

06. MODEL SUMMARY - IMPORTANT FEATURES





Using SHAP value to show the contribution or importance of each features

Reference: https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

Positives related:

- ❑ Gere_Drama: Genre of movie is drama
- ❑ OC_Uniter States: Original Country is US
- ❑ Actor_5: LDA actor topic #5
- ❑ Gere_Horror: Genre of movie is horror
- ❑ Director_5: LDA director topic #5

Negative related:

- ❑ Year: Publish year
- ❑ Gere_Documentary: Genre is documentary
- ❑ IG_English: Language is English
- ❑ Director_6: LDA director topic #6

06. Model Summary - Summary

Attribute sets: Basis + LDA

Baseline model [**OLS Linear** Regression]

For training set:

- ❑ Adjusted R-square: 0.40;

For testing set:

- ❑ MSE: 2.79*e+18

Best **Non-linear** model [Gradient Boosting]

For training set:

- ❑ Adjusted R-square: 0.92;

For testing set:

- ❑ MSE: 0.5656

Conclusion: Adding new latent factors (LDA Similarity) and using non-linear model (Gradient Boosting) can increase the predicting performance significantly.

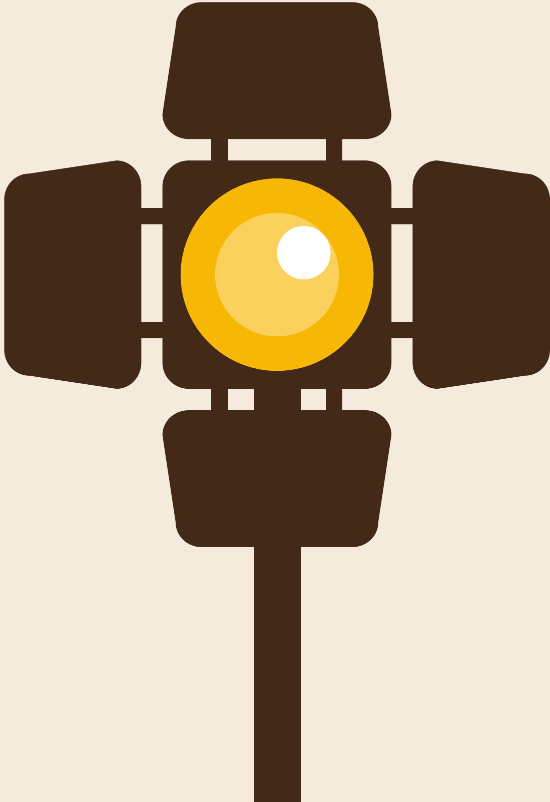
06. Model Summary - Evaluation

Predicted by best regressor

Prediction	***Actual Value***
5.9008	*5.9*
6.9009	*6.9*
6.4014	*5.9*
5.6984	*5.9*
7.3982	*5.9*

07. FUTURE WORKS

- Scrap more film data from different sources to get ‘genetal’ rating
  - Rotten tomatoes
  - Douban
  - ……
- Discover and refinine more features
  - Investment of film
  - Releasing on a holiday
  - ……
- Try more possible models:
  - word embedding
  - Xgboost regression
  - ……



8. REFERENCE

“Sort by Popularity - Most Popular Feature Films - IMDb.” *IMDb*, www.imdb.com/search/keyword/?title_type=movie. Accessed 20 Nov. 2022.

“Unigram Tokenization - Hugging Face Course.” *Unigram Tokenization - Hugging Face,* https://huggingface.co/course/chapter6/7?fw=pt. Accessed 20 Nov. 2022.

Kapadia, Shashank. “Evaluate Topic Models: Latent Dirichlet Allocation (LDA).” *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*, Towards Data Science， https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda. Accessed 20 Nov. 2022.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, Jiawei Han, "Automated Phrase Mining from Massive Text Corpora", accepted by IEEE Transactions on Knowledge and Data Engineering, Feb. 2018.

Jialu Liu*, Jingbo Shang*, Chi Wang, Xiang Ren and Jiawei Han, "Mining Quality Phrases from Massive Text Corpora”, Proc. of 2015 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'15), Melbourne, Australia, May 2015. (* equally contributed, slides)

Kelechava, Marc. *Using LDA Topic Models as a Classification Model Input*. https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28.

“Sentence-Transformers/NLI-Roberta-Large · Hugging Face.” *Sentence-Transformers/Nli-Roberta-Large · Hugging Face*, https://huggingface.co/sentence-transformers/nli-roberta-large.

THANK YOU

GITHUB REPOSITORY OF PROJECT

https://github.com/humphreyhuu/IMDb-FilmRating-NLP-Analysis.git