

IND ENG 242 Project Presentation: Predicting Box Office Revenue



Souhail Bantaleb, Romain Kakko-Chiloff, Yuyang Pan 雨阳, Alexandre
Vincent



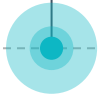
About

Using **web scraped data from IMDB** to
predict a movie's:

Gross Revenue
Opening Weekend Box Office
IMDB Rating

1. Timeline

**Web
Scraping**



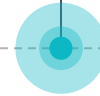
**Data
Cleaning
& Profiling**

**Feature
Selection**



**Machine Learning
Prediction**

**Model
Comparison**





2. Data

Data Collection : Web Scraping

- From IMDb website
- Web Scraping in Python using BeautifulSoup and Scrapy libraries
- Collect 25~informations about each 5000~
- Merge these data into a Dataset

IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

Enjoy unlimited streaming on Prime Video
Includes thousands of titles. Plans starting at \$8.99/mo [Start your 30-day free trial](#)

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE SHARE

Avatar (2009) 7.8 984 169 Rate This

Tous publics | 2h 42min | Action, Adventure, Fantasy | 16 December 2009 (France)

FROM THE DIRECTOR OF TERMINATOR 2 AND TITANIC

3:36 | Trailer 18 VIDEOS | 258 IMAGES

Watch Now
From \$2.99 (SD) on Prime Video ON TV ON DISC ALL

A paraplegic marine dispatched to the moon Pandora on a unique mission becomes torn between following his orders and protecting the world he feels is his home.

Director: [James Cameron](#)
Writer: [James Cameron](#)
Stars: [Sam Worthington](#), [Zoe Saldana](#), [Sigourney Weaver](#) | [See full cast & crew](#) »

- ## Gross

```

: final_data['Gross']
: 0      [' $2,787,965,087, ', '13 February 2015', '\n ...
1      []
2      [' $963,420,425, ', '25 November 2011', '\n ...
3      [' $880,674,175\n ' ]
4      [' $1,084,939,099\n ' ]
5      [' $260,502,115, ', '9 November 2013', '\n ...
6      [' $284,139,100\n ' ]
7      [' $591,794,936\n ' ]
8      [' $554,341,323\n ' ]
9      [' $1,405,413,868\n ' ]
10     [' $1,153,304,495\n ' ]
11     [' $873,260,194\n ' ]
12     [' $1,021,103,568\n ' ]
13     [' $934,416,487, ', '10 November 2011', '\n ...
14     [' $960,366,855, ', '9 January 2015', '\n ...
15     [' $956,019,788\n ' ]

```

Data Preprocessing

"Actor_1_id"	"Actor_1_name"	"Actor_2_id"	"Actor_2_name"
"Actor_3_id"	"Actor_3_name"	"Country"	"Director_id"
"Director_name"	"Genres"	"IMDb_critics"	"Keywords"
"Language"	"Movie_name"	"Storyline"	"Year"
"label_country"	"label_language"	"isAction"	"isAdventure"
"isAnimation"	"isBiography"	"isComedy"	"isCrime"
"isFantasy"	"isDocumentary"	"isFamily"	"isDrama"
"isHistory"	"isHorror"	"isMusic"	"isMystery"
"isRomance"	"isSciFi"	"isThriller"	"isSport"
"isSuperhero"	"isWar"	"isWestern"	"Duration_movie_final"
"Bugdet_final"	"Gross_final"	"Opening.Weekend.final"	"Actor_1_like_final"
"Actor_2_like_final"	"Actor_3_like_final"	"Director_like_final"	"IMDb_rating_final"

1. Remove null values in **Gross_final** - **2358** rows left

2. Add a column **"Successful"**:

Gross_final >= 2 * Budget_final	1
Gross_final < 2 * Budget_final	0

1522 rows

836 rows

3. Replace 0 in **Budget_final** with mean value

Data Profiling: One-Hot Encoding

Country/ Region	Europe	North America	Oceania	Asia	Africa	Other
	0	1	2	3	4	5

Language	English	Japanese	French	Mandarin	Spanish	Hindi	German	Other
	0	1	2	3	4	5	6	7

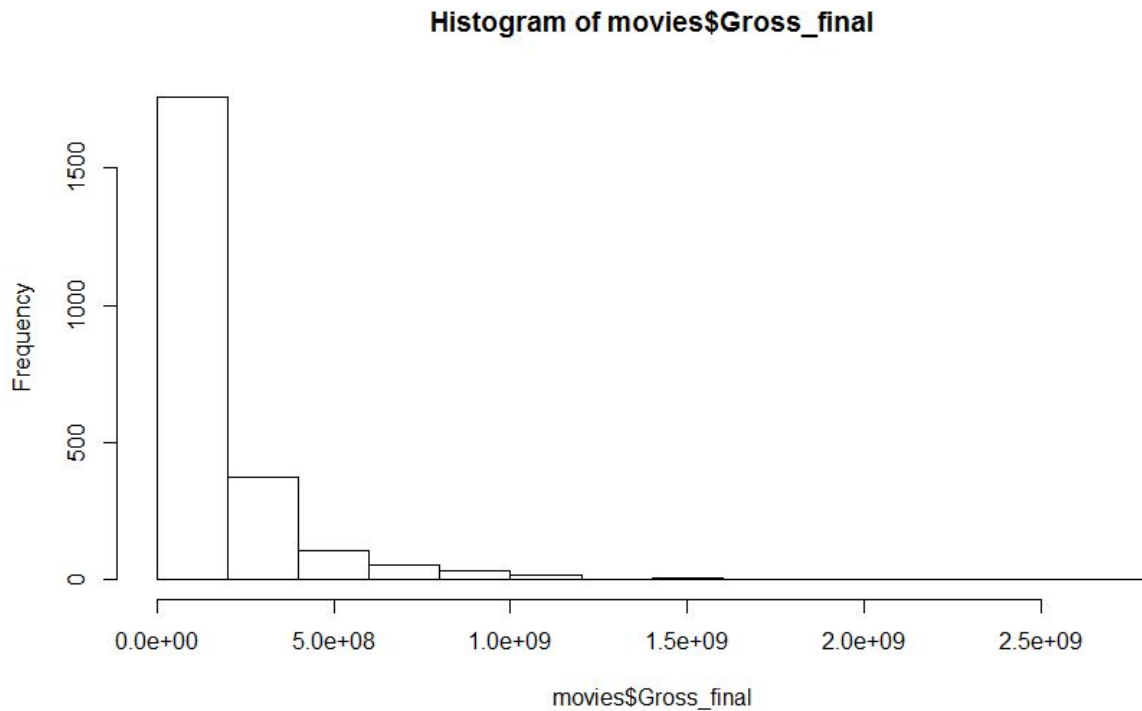
	isAction	isAdventure	isAnimation	...	isWestern
Action	1	0	0		0
Adventure	0	1	0		0
Animation	0	0	1		0
...	0	0	0		0
Western	0	0	0		1

Genre
Country/Region
Language

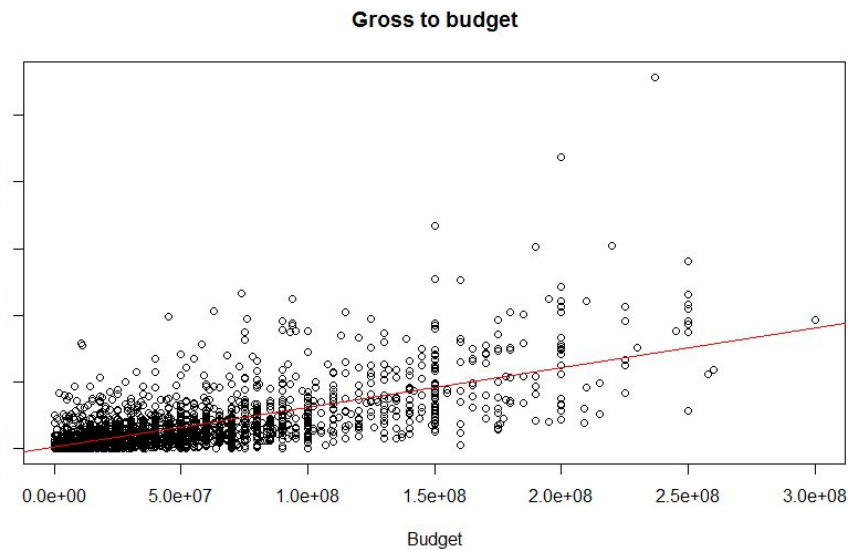
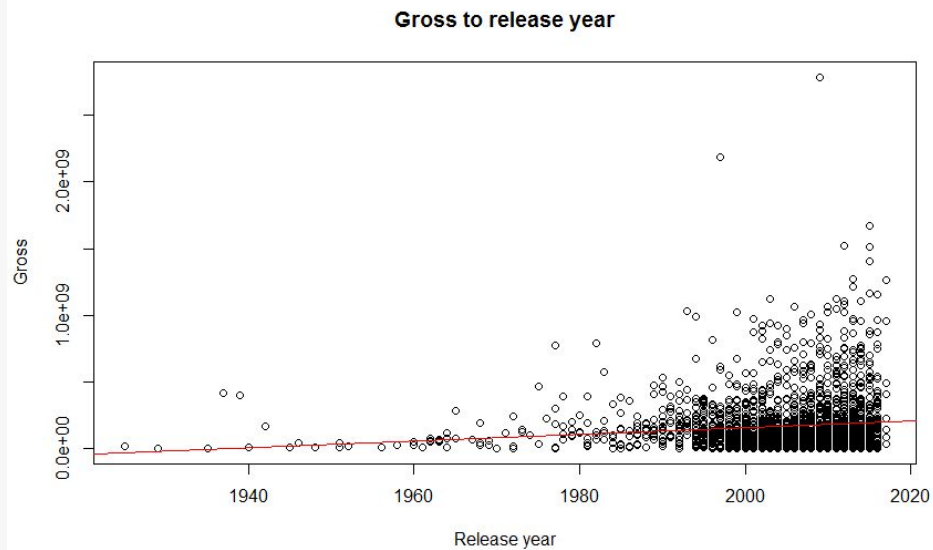
1

0

Data Exploration Analysis

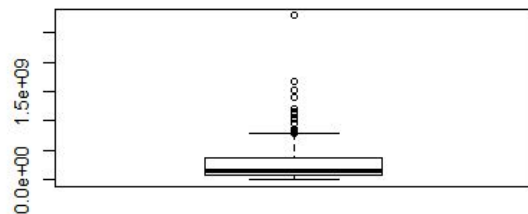


Data Exploration Analysis:

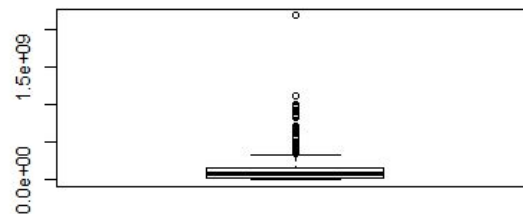


Data Exploration Analysis

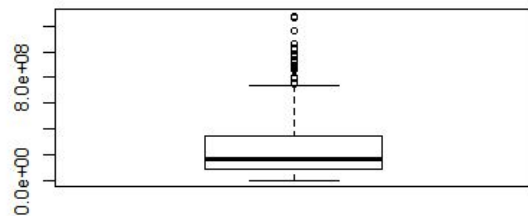
SciFi



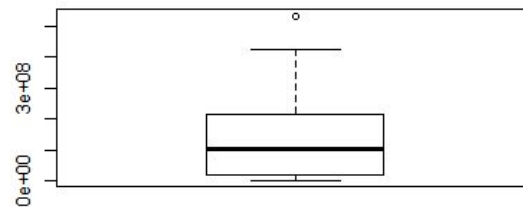
Drama



Family



Western

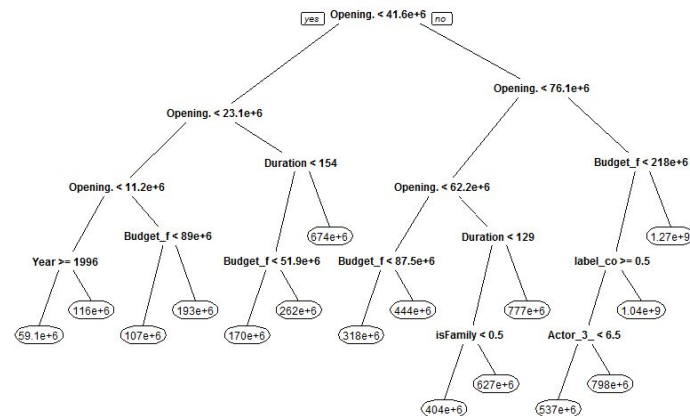


A background image of a man in a Roman-style military uniform, including a laurel wreath and a chain of office, sticking his tongue out. The image is overlaid with a teal gradient.

3. Analytics Models

Linear Regression, Logistic Regression, LDA, CART, Random Forest, Boosting, NLP

Model	Score	Type
Linear Regression	Accuracy = 55.8%	Linear
Boosting	CV R ² = 0.488	
Random Forest	OSR2 = 0.73	
Logistic Regression	Accuracy = 54.85% AUC = 0.65	Classification
LDA	Accuracy = 66% AUC = 0.68	
CART	OSR2 = 0.59	
NLP	Accuracy = 67%	



15 NLP to predict the success of a movie

Storyline

[Edit](#)

After Elizabeth, Will, and Captain Barbossa rescue Captain Jack Sparrow from the the land of the dead, they must face their foes, Davy Jones and Lord Cutler Beckett. Beckett, now with control of Jones' heart, forms a dark alliance with him in order to rule the seas and wipe out the last of the Pirates. Now, Jack, Barbossa, Will, Elizabeth, Tia Dalma, and crew must call the Pirate Lords from the four corners of the globe, including the infamous Sao Feng, to gathering. The Pirate Lords want to release the goddess Calypso, Davy Jones's damned lover, from the trap they sent her to out of fear, in which the Pirate Lords must combine the 9 pieces that bound her by ritual to undo it and release her in hopes that she will help them fight. With this, all pirates will stand together and will make their final stand for freedom against Beckett, Jones, Norrington, the Flying Dutchman, and the entire East India Trading Company.

Written by J. Curcio

[Plot Summary](#) | [Plot Synopsis](#)

1 : 
Success

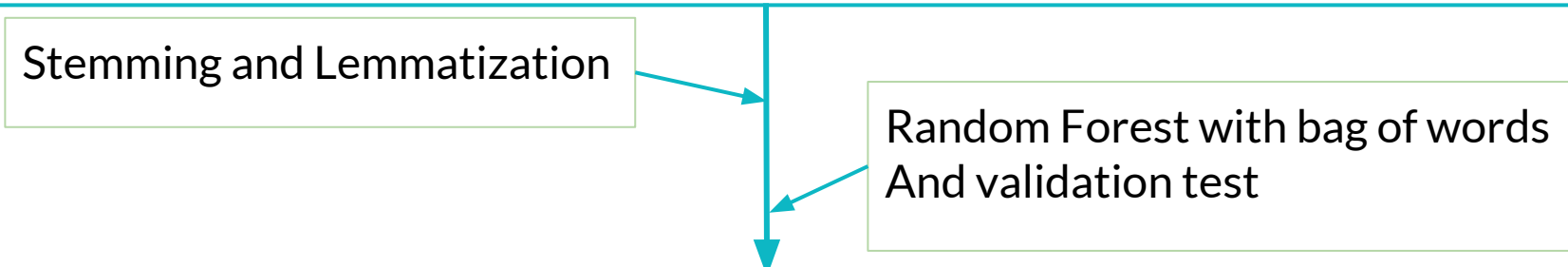
0 : Not a
success

Cleaning storylines to then use bag of words

'elizabeth captain barbossa rescue captain jack sparrow land dead must face foes davy jones lord c
utler beckett beckett control jones heart forms dark alliance order rule seas wipe last pirates ja
ck barbossa elizabeth tia dalma crew must call pirate lords four corners globe including infamous
sao feng gathering pirate lords want release goddess calypso davy jones damned lover trap sent fea
r pirate lords must combine pieces bound ritual undo release hopes help fight pirates stand togeth
er make final stand freedom beckett jones norrington flying dutchman entire east india trading com
pany'

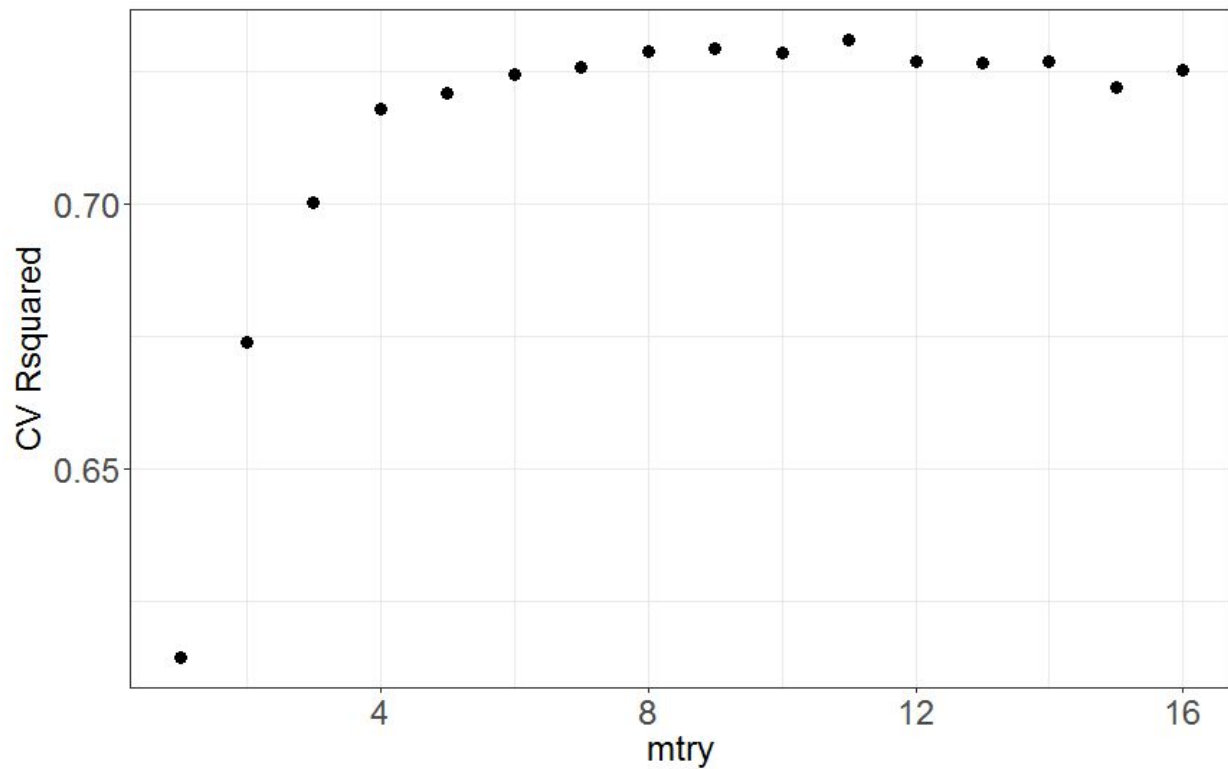
Stemming and Lemmatization

Random Forest with bag of words
And validation test



	Predicted 0	Predicted 1
True Label 0	14	151
True Label 1	9	297

Random forest: parameter selection



4. Impact

A man in a tactical vest is shown from the waist up, shouting with his mouth wide open and arms outstretched. He is wearing a black tactical vest with multiple straps and buckles. The background is a bright, hazy outdoor setting. The image is overlaid with a teal gradient on the left side where the text is located.

A movie's success depends on several factors ie. cast, budget, release time, ...

Can we predict a movie's

1. **Revenue**
2. **Opening weekend box office**
3. **IMDB rating**

Before its release date?

By building a model that could help predict the expected revenue of a movie, producers and studios might be able to make better decisions, for instance when allocating resources.

Movie Studios & Producers



- Estimate box office revenue
 - Make a better decision on choosing:
 - **Genre**
 - **Actors**
 - **Directors**
 - **Duration of movie**
- To maximize revenue

1. Model Accuracy: Classifier Parameters

- ▷ Tune the random forest parameters:

testing out multiple minimum sample leaf sizes

2. Model Accuracy: Adding Features**3. Building more complex models****4. Predict Opening weekend box office & IMDB ratings****5. Testing: New Releases (Opening Weekend Box Office)**



Thank You