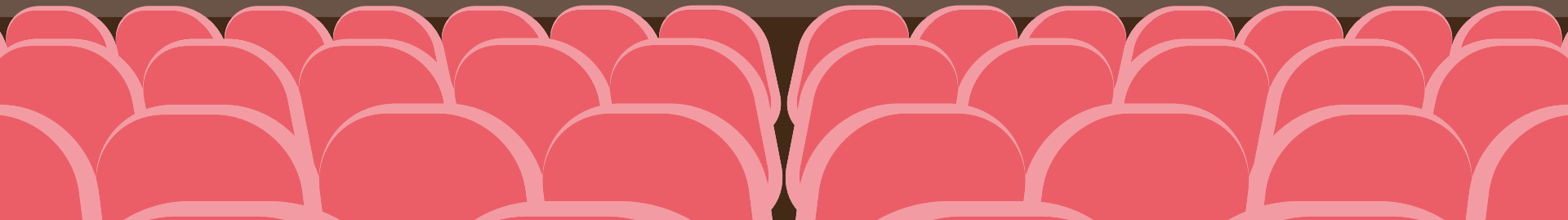
MOVIE INVESTMENT

DSI 14 - Cecilia Oliveira

Capstone Project

2nd Dec 2020



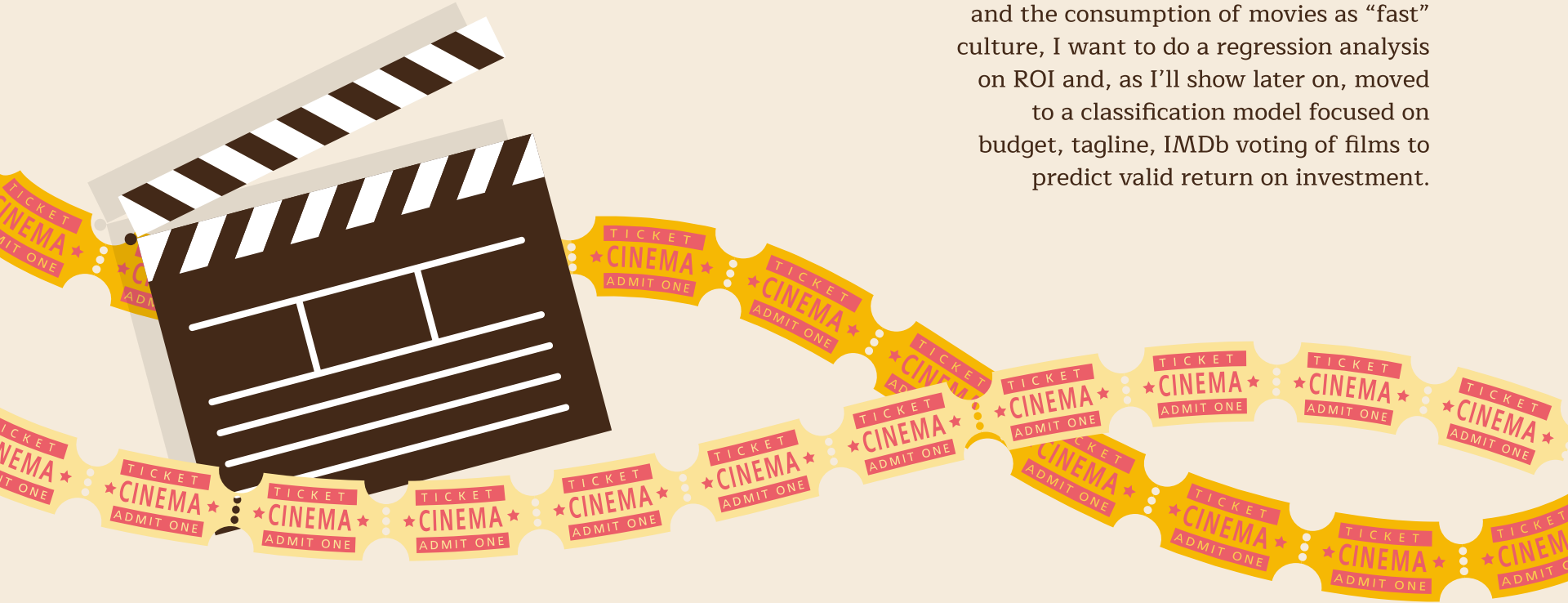
AGENDA

1. BUSINESS CASE
2. THE DATA
   And Features
3. PRE-PROCESSING AND VARIABLES
4. EDA
5. MODEL & RESULTS
6. FOR FUTURE EXPLORATION

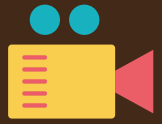
BUSINESS CASE

BUSINESS CASE

With the increase of streaming platforms and the consumption of movies as “fast” culture, I want to do a regression analysis on ROI and, as I’ll show later on, moved to a classification model focused on budget, tagline, IMDb voting of films to predict valid return on investment.

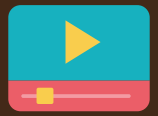


BUSINESS CASE



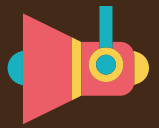
Potential Audience

Cultural Investors, Corporates, Production Companies, Government, Film Councils and Festivals



Goal

Increase Popularity, better investment decisions based on budget, genre, tagline ...



Success Metrics

ROI, Revenue, IMdb votes, investment validity

THE DATA

And features

THE DATA

Shape : (45466, 24)

Kaggle - https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv

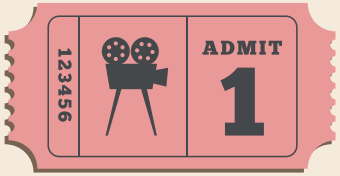
Column_Name	DataType	Measurement_Unit	Description
adult	object	NA	Adult movie
belongs_to_collection	object	NA	Film franchise
budget	object	USD	Budget at production
genres	object	NA	Movie genre
homepage	object	NA	Website
id	object	NA	Identification number
imdb_id	object	NA	Id @ IMDb.com
original_language	object	NA	Movie language
original_title	object	NA	Title in original language
overview	object	NA	Short movie description
popularity	object	NA	Popularity score
poster_path	object	NA	Movie poster jpg
production_companies	object	NA	Film production company
production_countries	object	NA	Company country
release_date	object	NA	Cinema date release
revenue	float64	USD	Revenue generated
runtime	float64	minutes	Movie duration
spoken_languages	object	NA	Languages spoken
status	object	Released, Rumored, ...	Movie status
tagline	object	NA	One line description
title	object	NA	Movie official title
video	object	NA	Movie website
vote_average	float64	NA	Average movie score
vote_count	float64	NA	Number of votes

PRE-PROCESSING AND VARIABLES

PRE-PROCESSING & NEW VARIABLES

production_companies	production_countries
[{'name': 'Pixar Animation Studios', 'id': 3}]	[{'iso_3166_1': 'US', 'name': 'United States o...
[{'name': 'TriStar Pictures', 'id': 559}, {'na...	[{'iso_3166_1': 'US', 'name': 'United States o...

Columns review



Change Dtype

Drop & Delete

delete unwanted info & drop homepage, poster_path, video, id and imdb_id columns

No null!

Dropped null budgets & revenue

```
df.genres.unique()
```

```
array(['Animation', 'Adventure', 'Romance', 'Comedy', 'Action', 'Family',
       'History', 'Drama', 'Crime', 'Fantasy', 'Science Fiction', nan,
       'Thriller', 'Music', 'Horror', 'Documentary', 'Mystery', 'Western',
       'TV Movie', 'War', 'Foreign', 'Carousel Productions', 'Aniplex',
       'Odyssey Media'], dtype=object)
```

Fillna & Row Drop

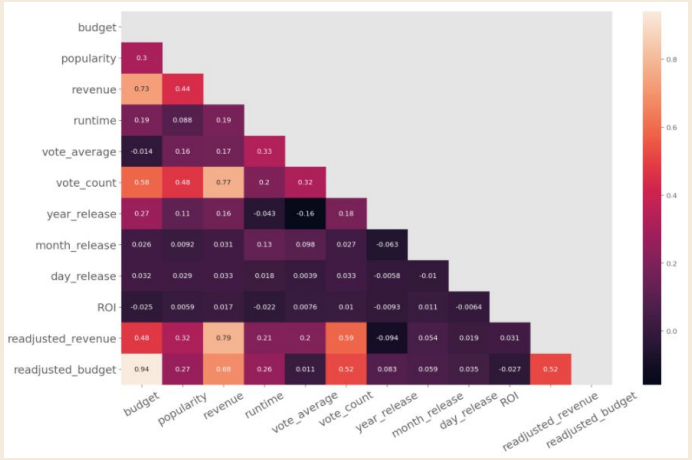
Used fillna & dropped columns without popularity, release_date

New Variables

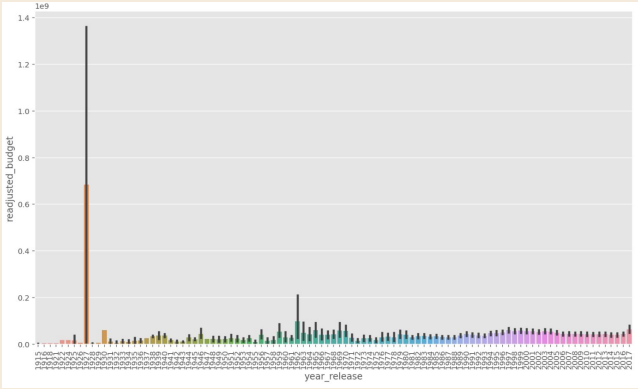
Split year, month and day release and created ROI.

EDA

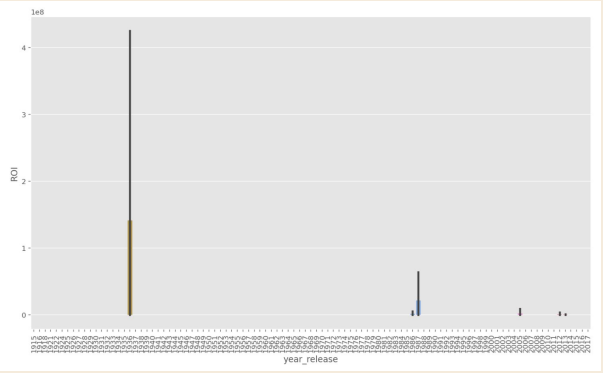
EDA



Revenue & Budget - Number of Votes & Scores
ROI - not as related



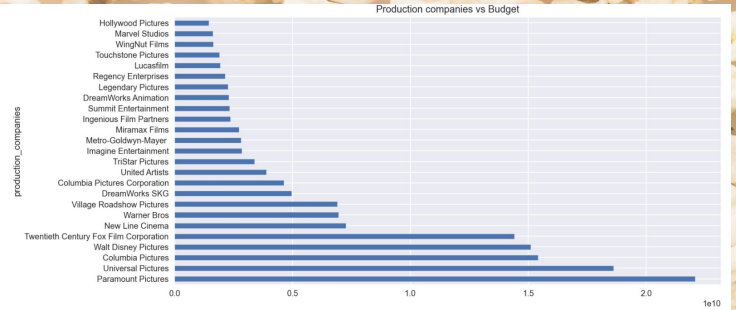
Steady budget increase
1927 - Metropolis - budget USD 92,620,000

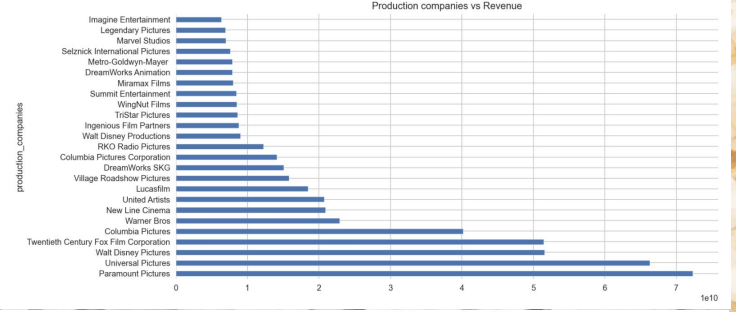


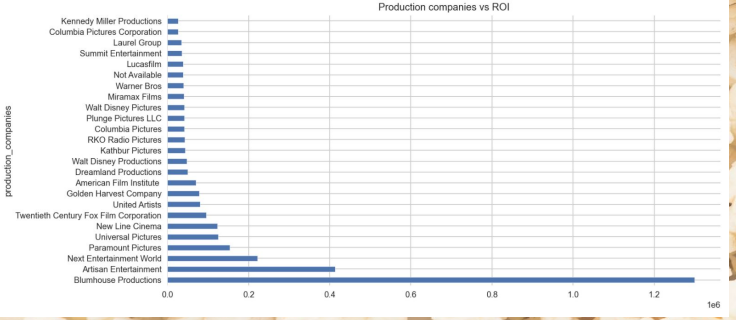
Outlier 1937 - Snow White & the Seven Dwarfs - 12,424% ROI

EDA

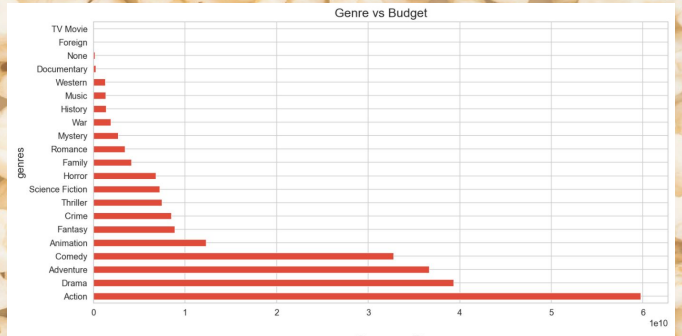
Production Companies

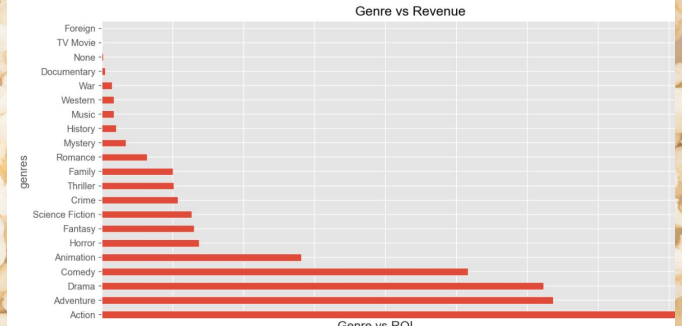


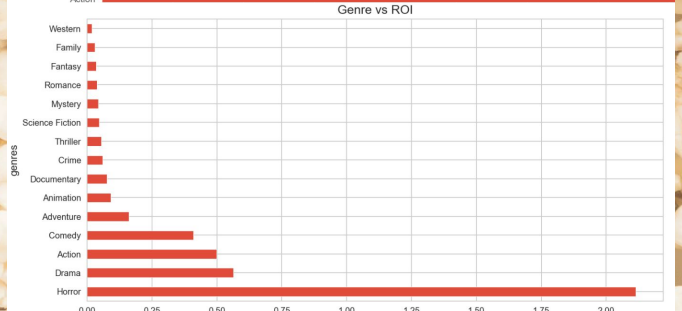




Genres

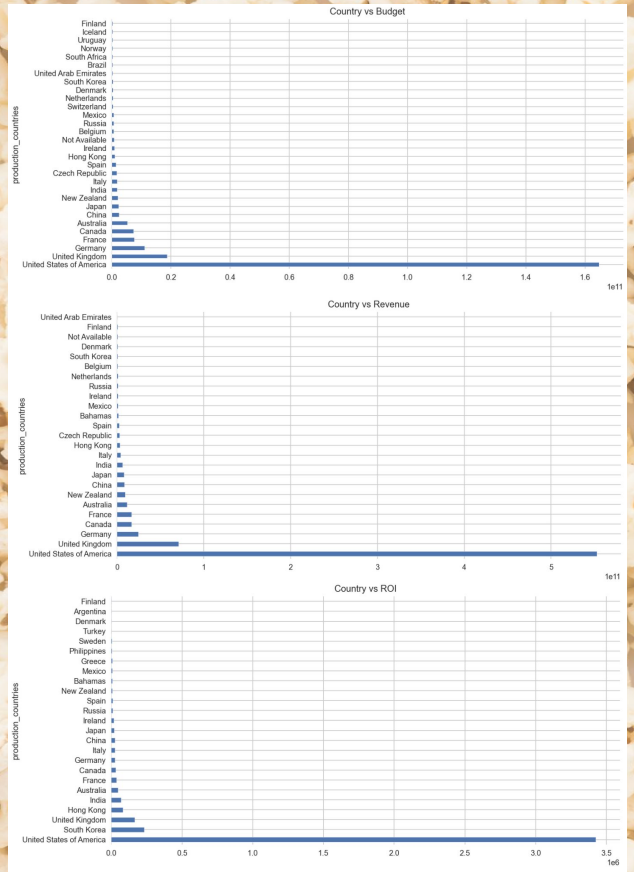






EDA

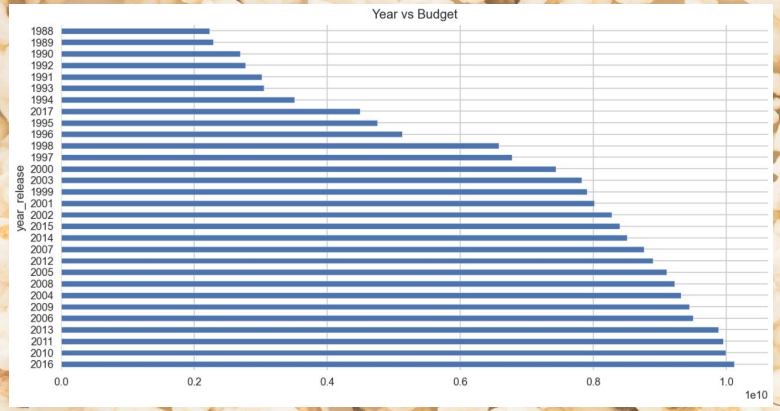
Production Countries

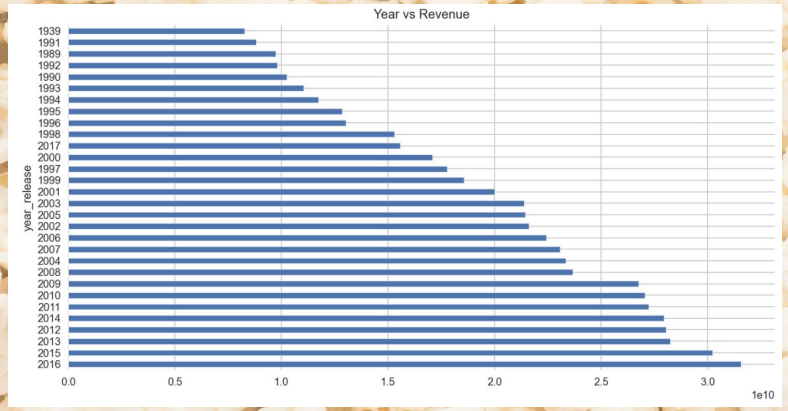


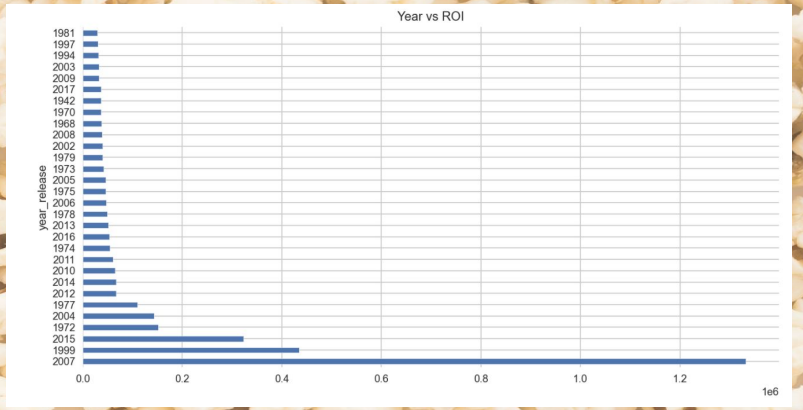
IMDb scores

Year Release

EDA





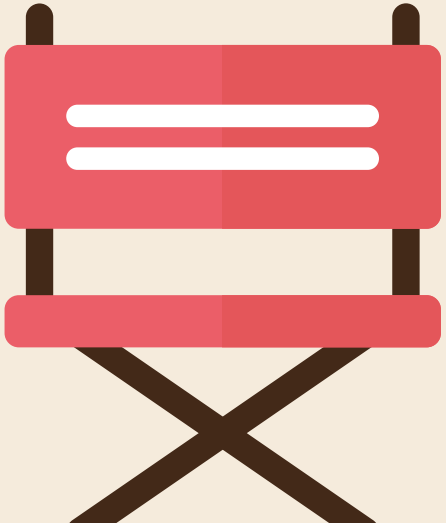


MODELS & RESULTS

THE MODELS

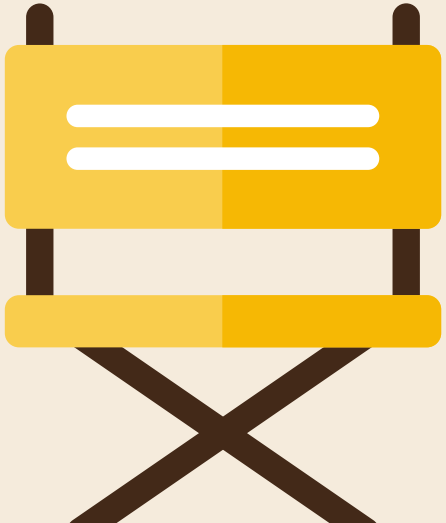
Regression

ROI, IMDb scores & Revenue



Classification

Valid_investment

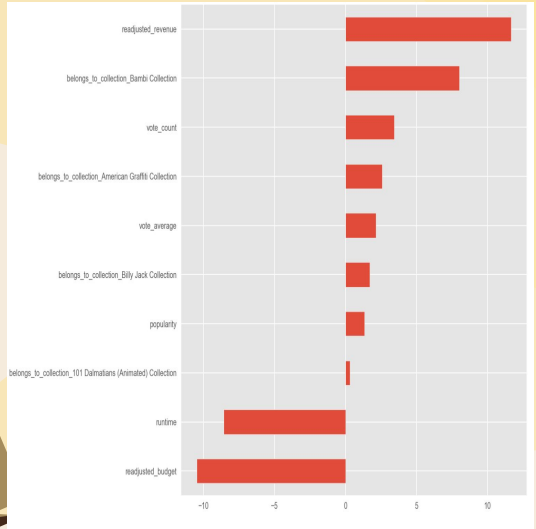


Natural Language Processing

THE MODELS REGRESSION

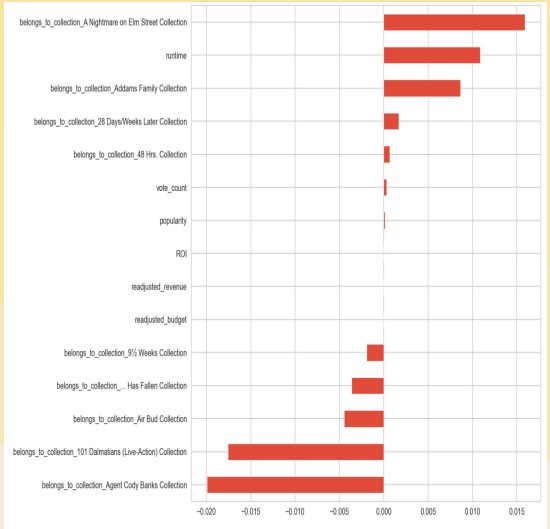
ROI

Train score: 0.9892423
Test score: 0.30540365
Cross-validated training scores: -1621860.726
Gradient Boosting Regressor



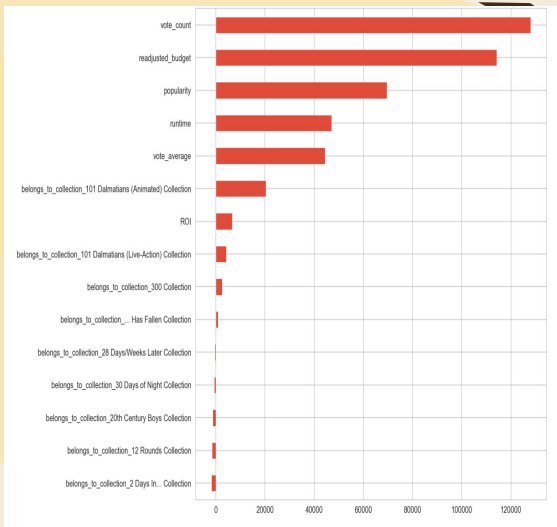
Vote Average

Training score: 0.61227065
Test Score: 0.3286498
Mean cross validation score: 0.094356950
RidgeCV

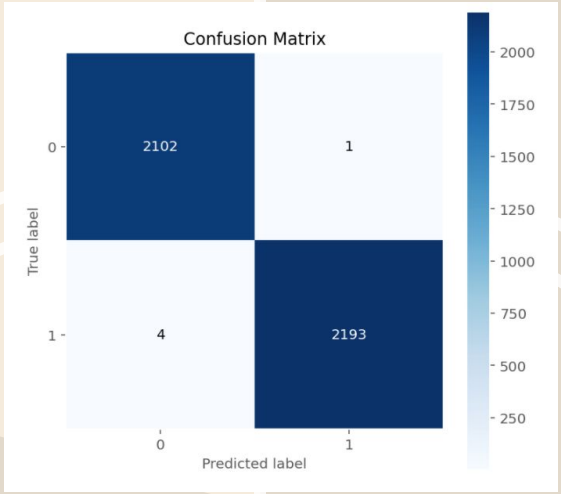


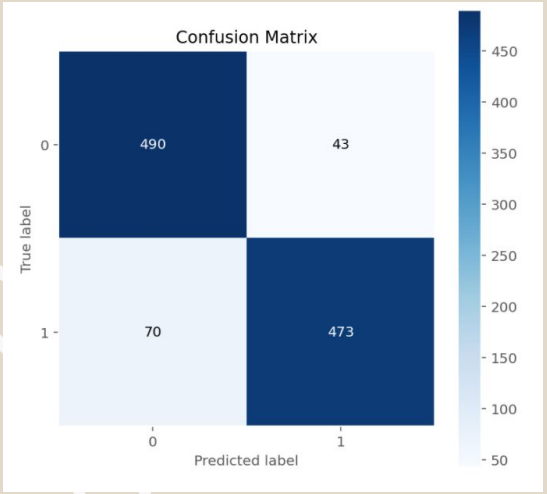
Revenue

Train score: 0.8809531
Test score: 0.57763
Cross-validated training scores: 0.788107
Gradient Boosting Regressor



THE MODELS CLASSIFICATION





Valid ROI (1) or not (0)

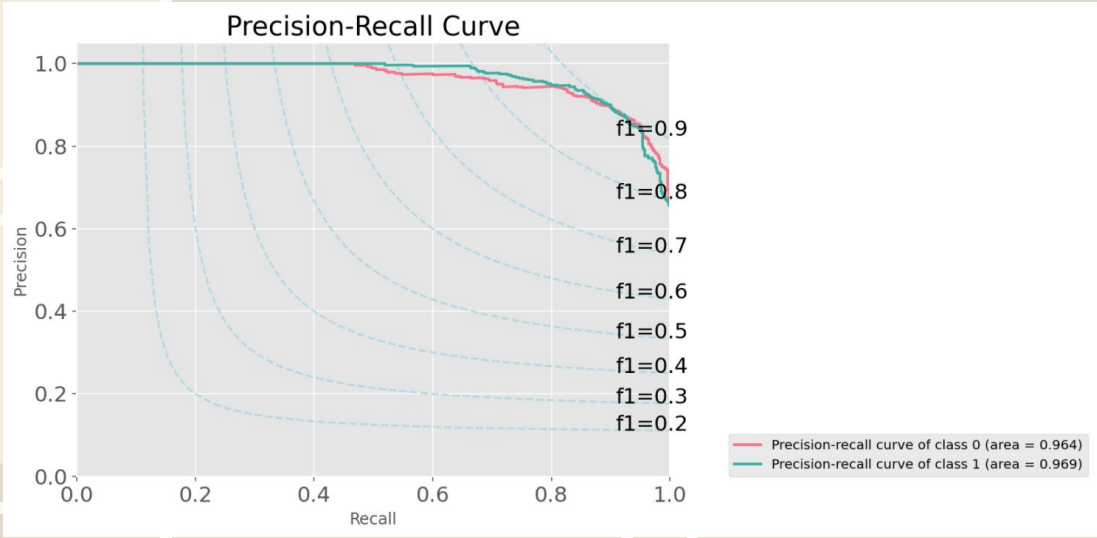
0 - not valid investment
1 - valid investment

Training score: 0.99883
Test Score: 0.894981
Best score: 0.8951162
Logistic Regression

TRAIN SET	PRECISION	RECALL	F1-SCORES	SUPPORT
0 /- Invst	1.00	1.00	1.00	2103
1/+ Invst	1.00	1.00	1.00	2197
Accuracy			1.00	4300
Macro Avg	1.00	1.00	1.00	4300
Weighted avg	1.00	1.00	1.00	4300

TEST SET	PRECISION	RECALL	F1-SCORES	SUPPORT
0 /- Invst	0.88	0.92	0.90	533
1/+ Invst	0.92	0.87	0.89	543
Accuracy			0.89	1076
Macro Avg	0.90	0.90	0.89	1076
Weighted avg	0.90	0.89	0.89	1076

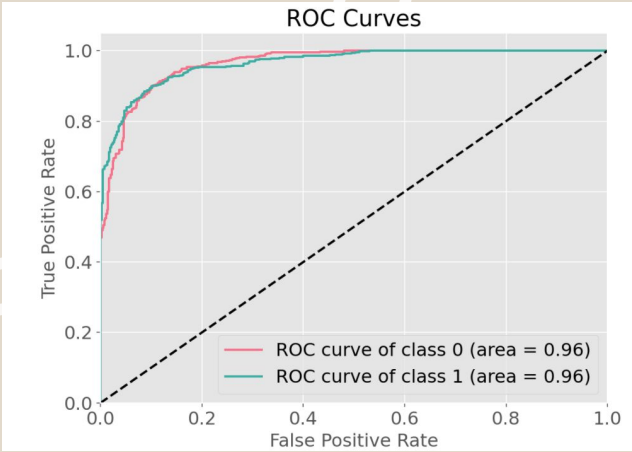
THE MODELS CLASSIFICATION



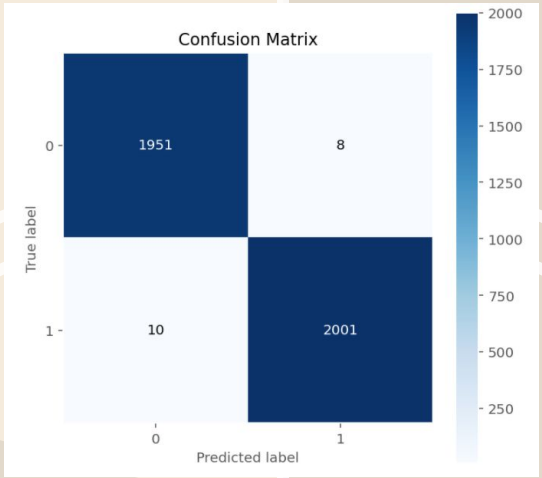
Valid ROI (1) or not (0)

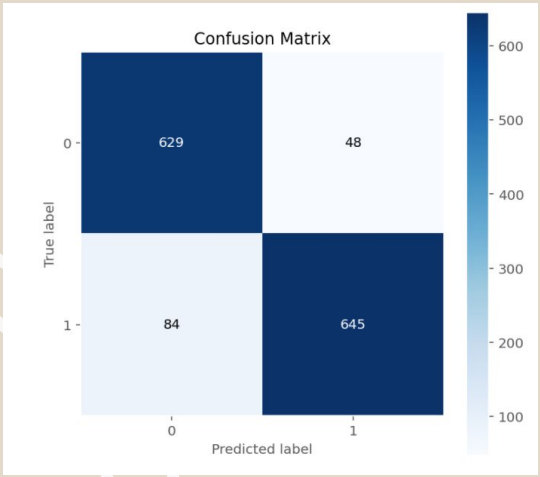
0 - not valid investment
1 - valid investment

Training score: 0.99883
Test Score: 0.894981
Best score: 0.8951162
Logistic Regression



THE MODELS CLASSIFICATION





Valid ROI (1) or not (0) pre/post 2011

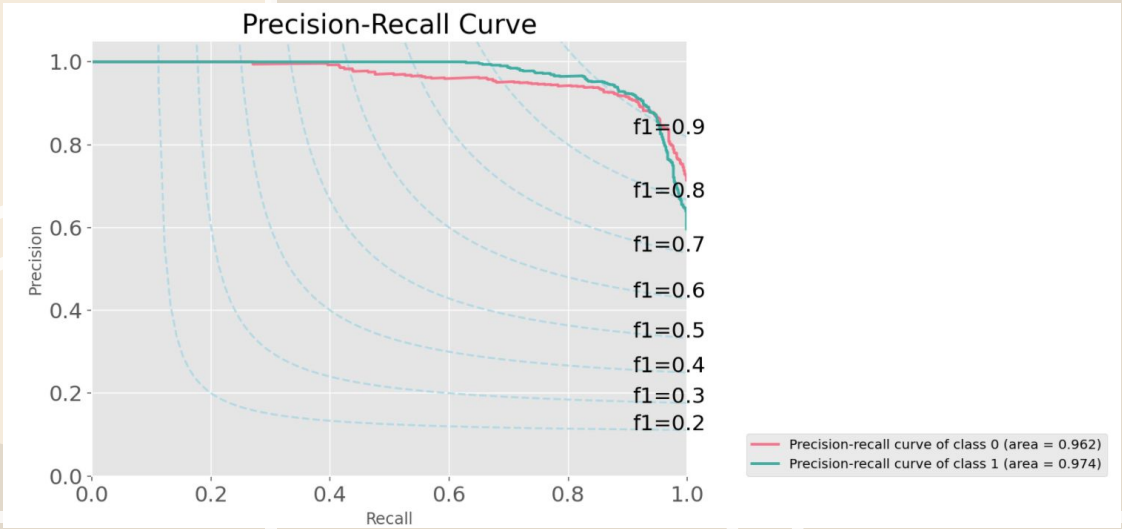
0 - not valid investment
1 - valid investment

Streaming Services launch
Train score: 0.99546599
Test score: 0.906116
Best Score: 0.93148
Logistic Regression

TRAIN SET	PRECISION	RECALL	F1-SCORES	SUPPORT
0/negative Inv	0.99	1.00	1.00	1959
1/positive I	1.00	1.00	1.00	2011
Accuracy			1.00	3970
Macro Avg	1.00	1.00	1.00	3970
Weighted avg	1.00	1.00	1.00	3970

TEST SET	PRECISION	RECALL	F1-SCORES	SUPPORT
0	0.88	0.93	0.91	677
1	0.93	0.88	0.91	729
Accuracy			0.91	1406
Macro Avg	0.91	0.91	0.91	1406
Weighted avg	0.91	0.91	0.91	1406

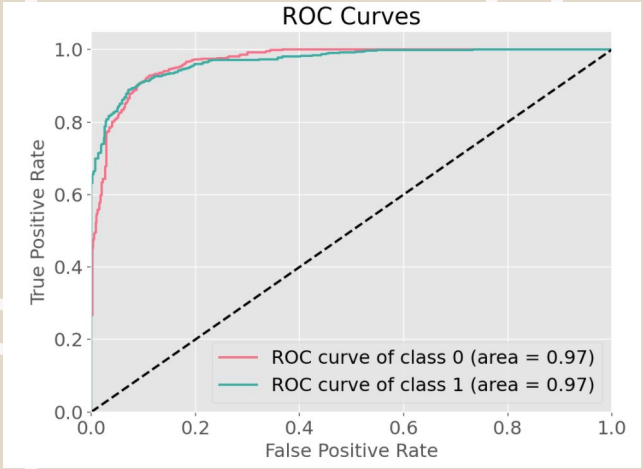
THE MODELS CLASSIFICATION



Valid ROI (1) or not (0) pre/post 2011

Streaming Services launch
0 - not valid investment
1 - valid investment

Train score: 0.99820971
Test score: 0.88872832
GridSearchCV/Logistic Regression

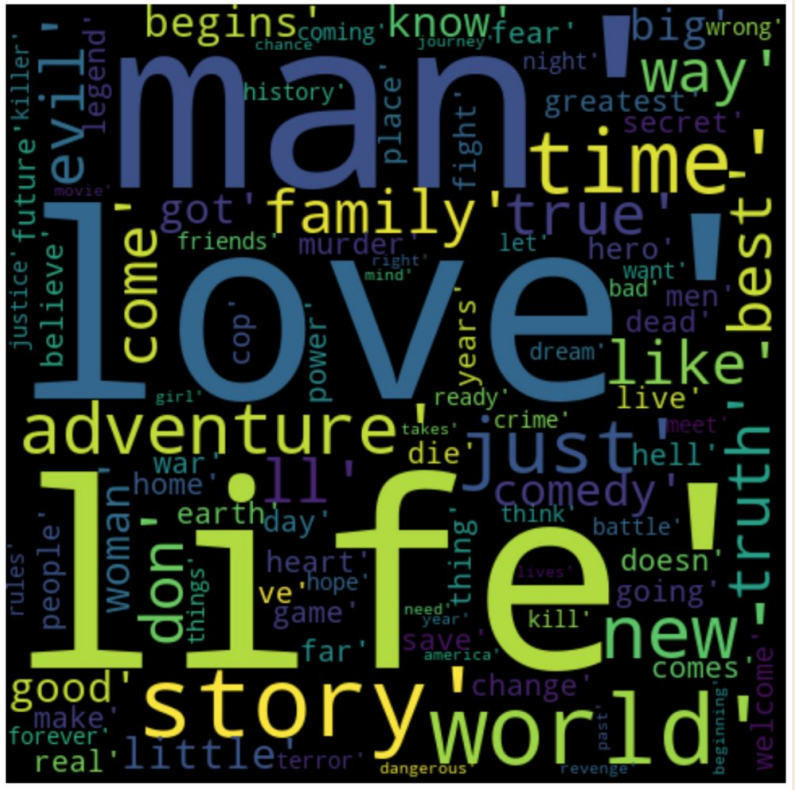


THE MODELS

NLP

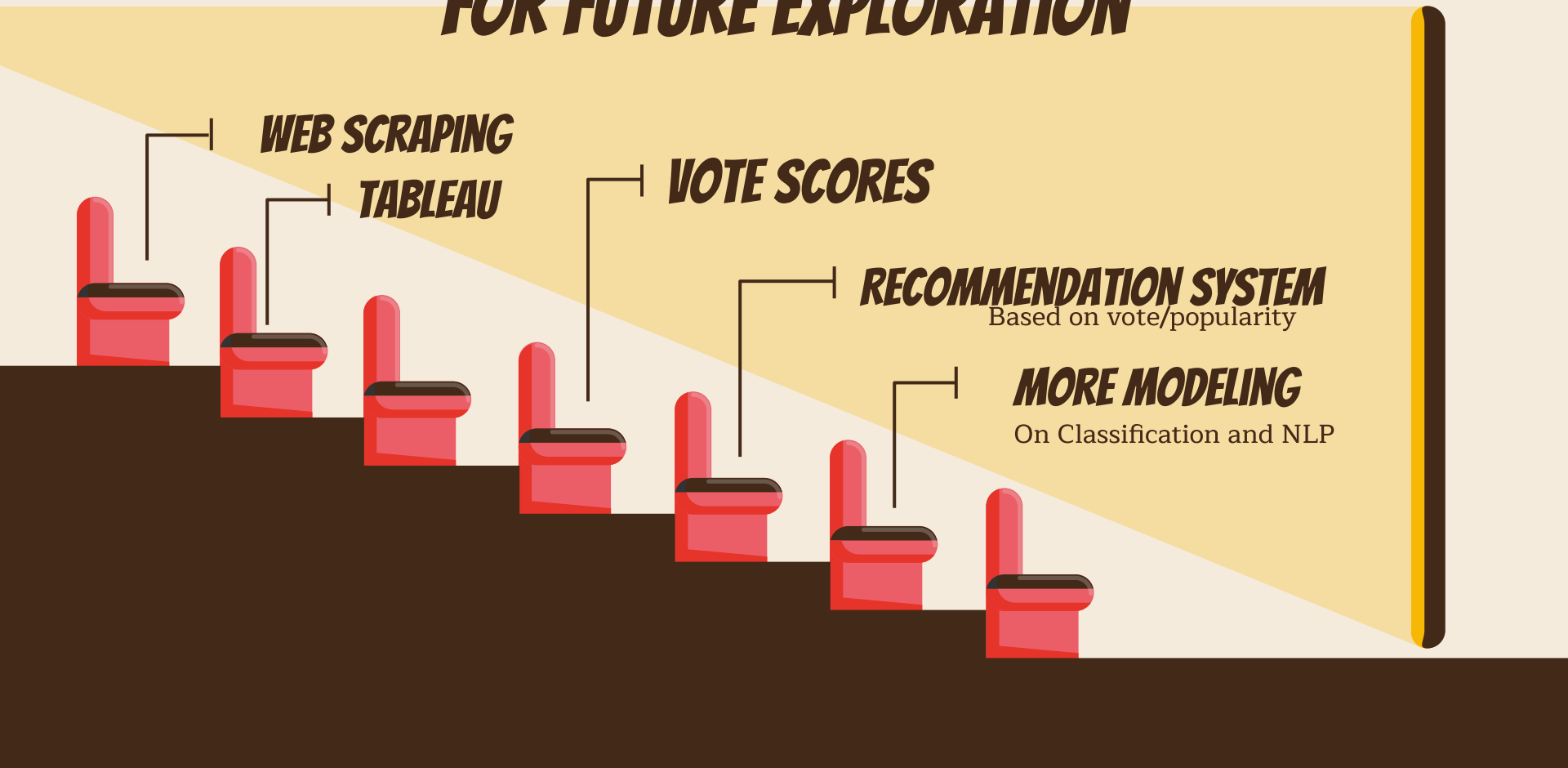
Best Scores

Training score: 1.0
Test Score: 1.0
Cross validation: 0.9994535519125683



FOR FUTURE EXPLORATION

FOR FUTURE EXPLORATION



THANKS !

QUESTIONS?

oliveira.ceci@gmail.com