THE PERFECT MOVIE RECIPE

Hien Lê
Zafer Kocaoglu
Francesco Maizza
Anita Mezzetti
Nataliia Surianinova

01 CINEMA INDUSTRY

02 DATA WRANGLING

03 FEATURE ENGINEERING

04 EXPLORATORY DATA ANALYSIS

05 THE MODELS

06 INTERPRETATION OF THE RESULTS

07 IMPLICATIONS FOR BUSINESSES

THE CINEMA INDUSTRY

01

TRENDS IN THE MOVIE INDUSTRY

The cinema industry has been in a continuous path of **INNOVATION** and **CHANGE**

Streaming revolution

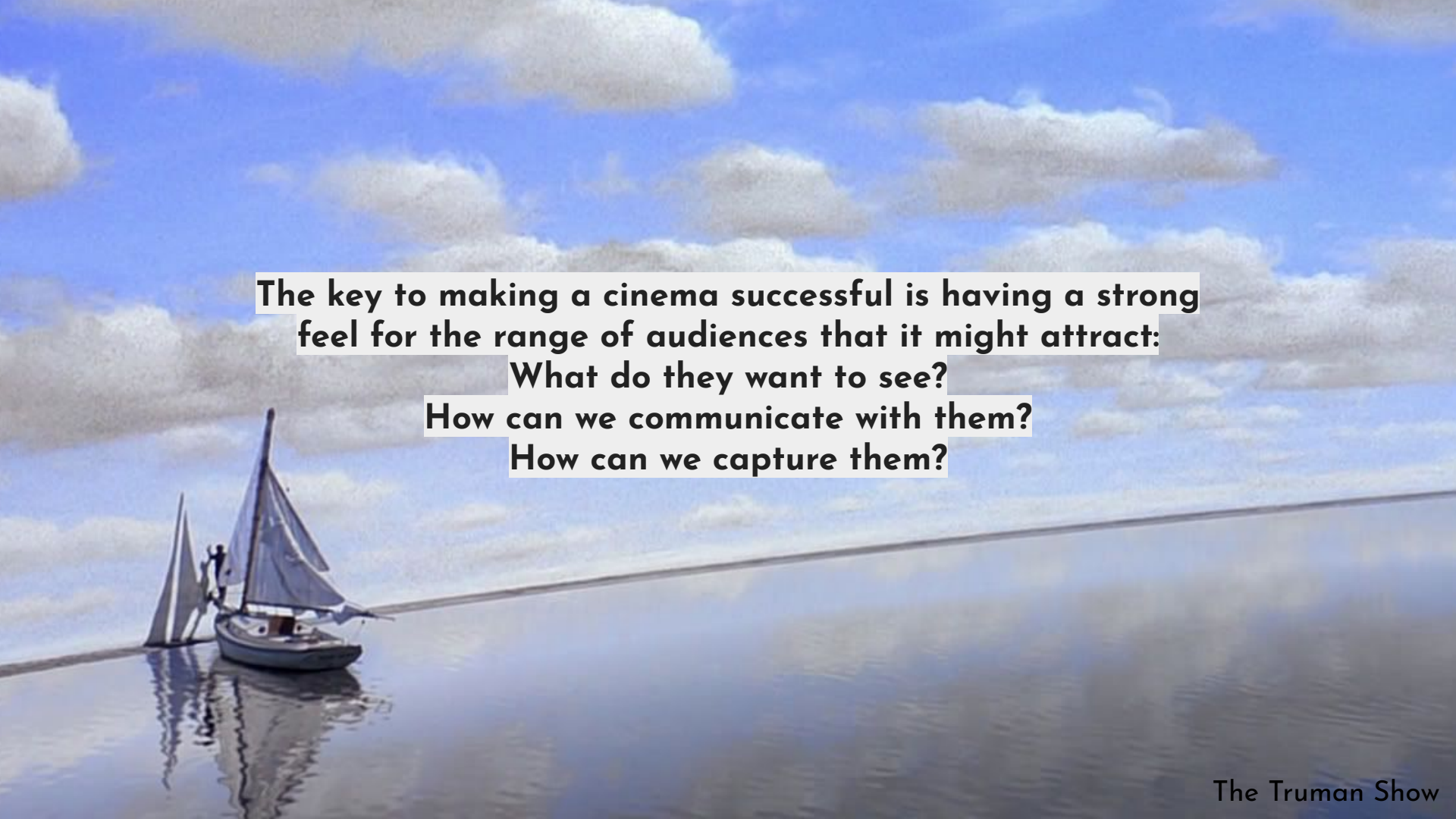
Social media influence

- Share opinions
- Promote
- Engaging experience
- Collect data

Blockbuster movie: highly popular and financially successful movie

Out of the top 30 most profitable films of the last 10 years, **Frozen** is the only original story

Stories are not new → Adapt the story line to a continuously evolving society



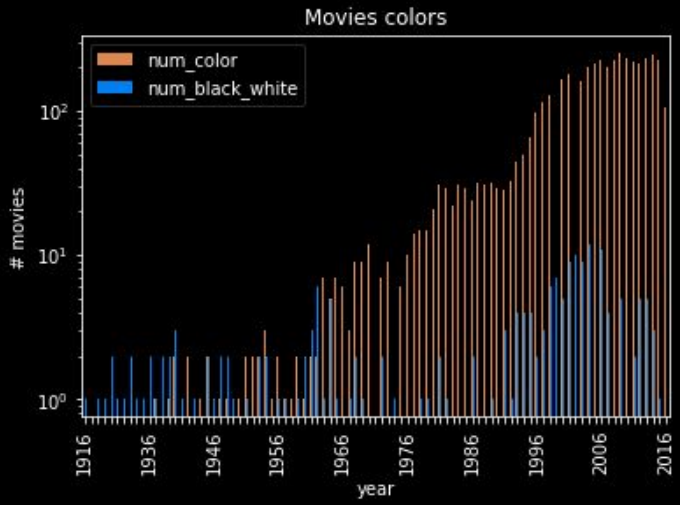
The Truman Show

DATA WRANGLING

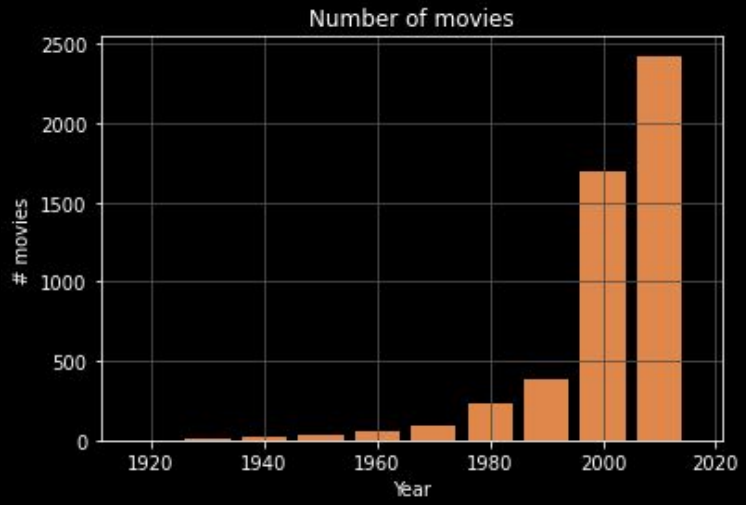
02

FEATURES ANALYSIS

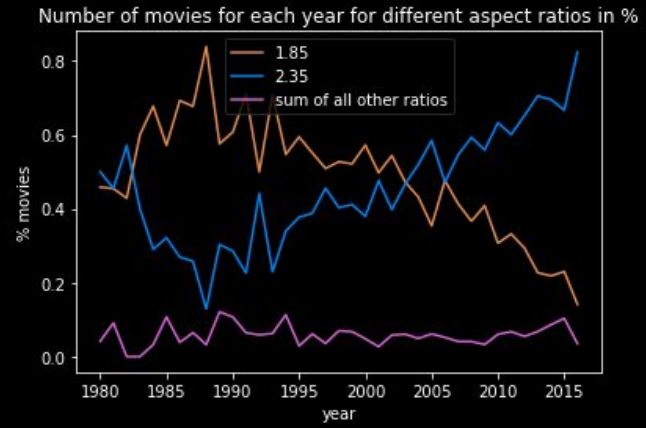
COLOR



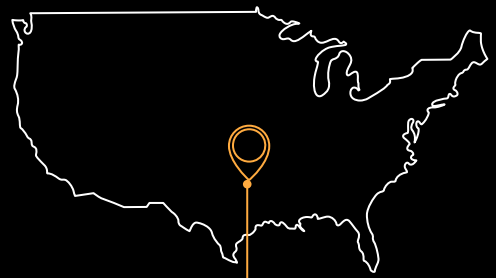
YEAR



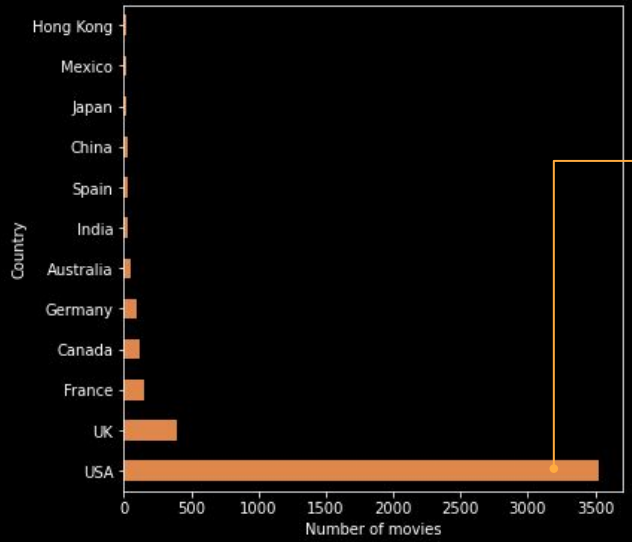
ASPECT RATIO



COUNTRY



USA



FILL NaN VALUES WITH DATA FROM OTHER DATASETS

WIKIPEDIA

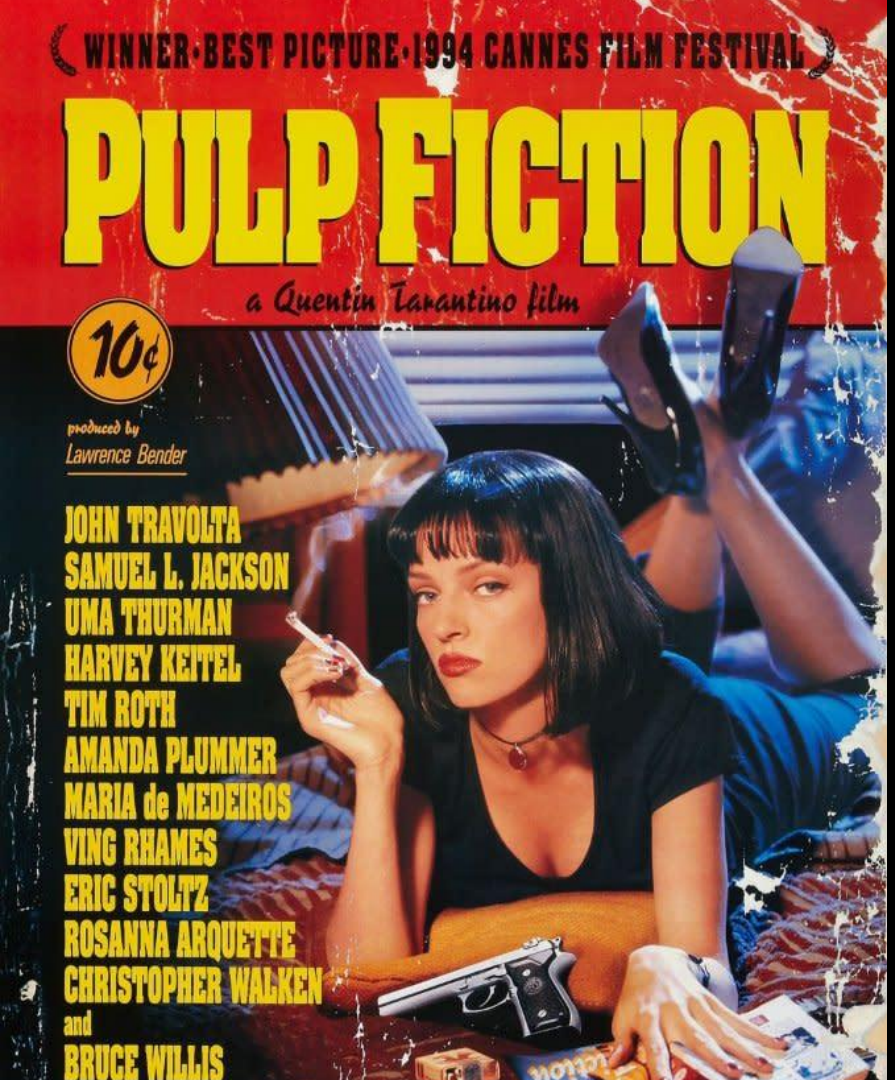
Gross
Budget

TMDB

Budget
Keywords

Missing values

	Before	After
Gross	604	577
Budget	369	171
Keywords	138	129

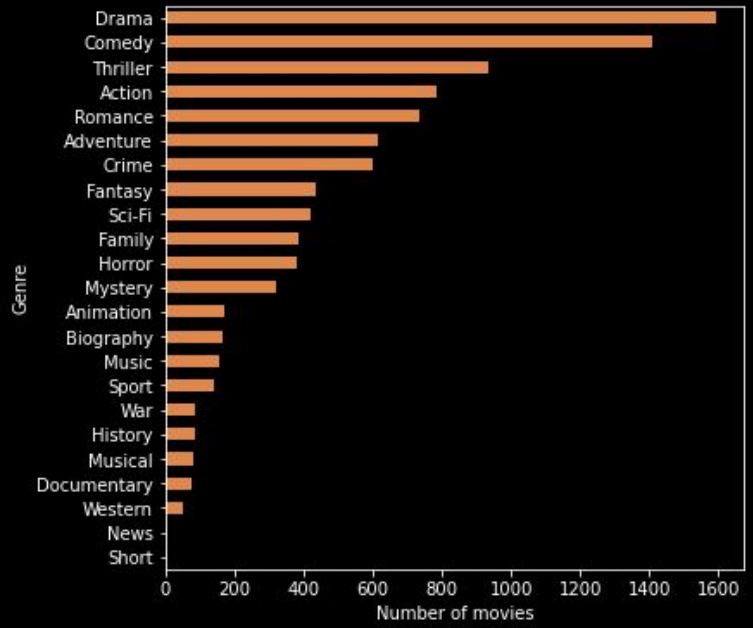


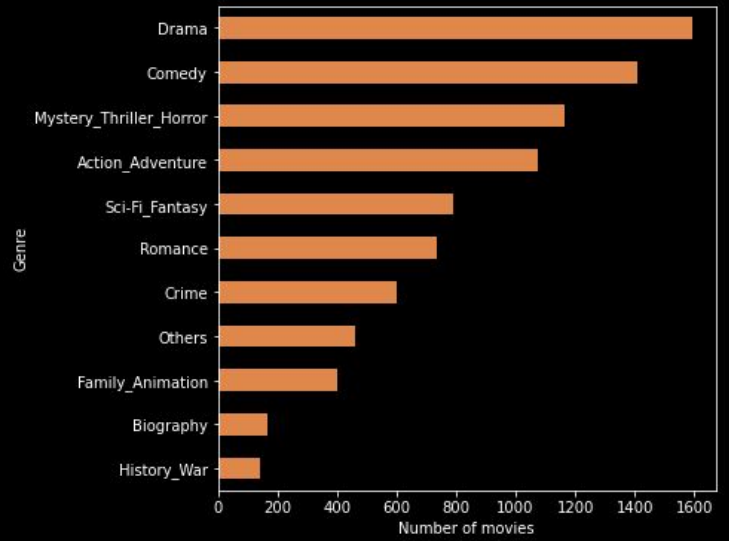
FEATURE ENGINEERING

03

GENRES

- Genres' number reduction
- One Hot Encoding





RANKING

DIRECTORS

- Number of Movies
- IMDB Score

Steven Spielberg

Clint Eastwood

Woody Allen

Martin Scorsese

ACTORS

- Number of Movies
- Facebook Likes

Actor 1			
Actor 2			
Actor 3			

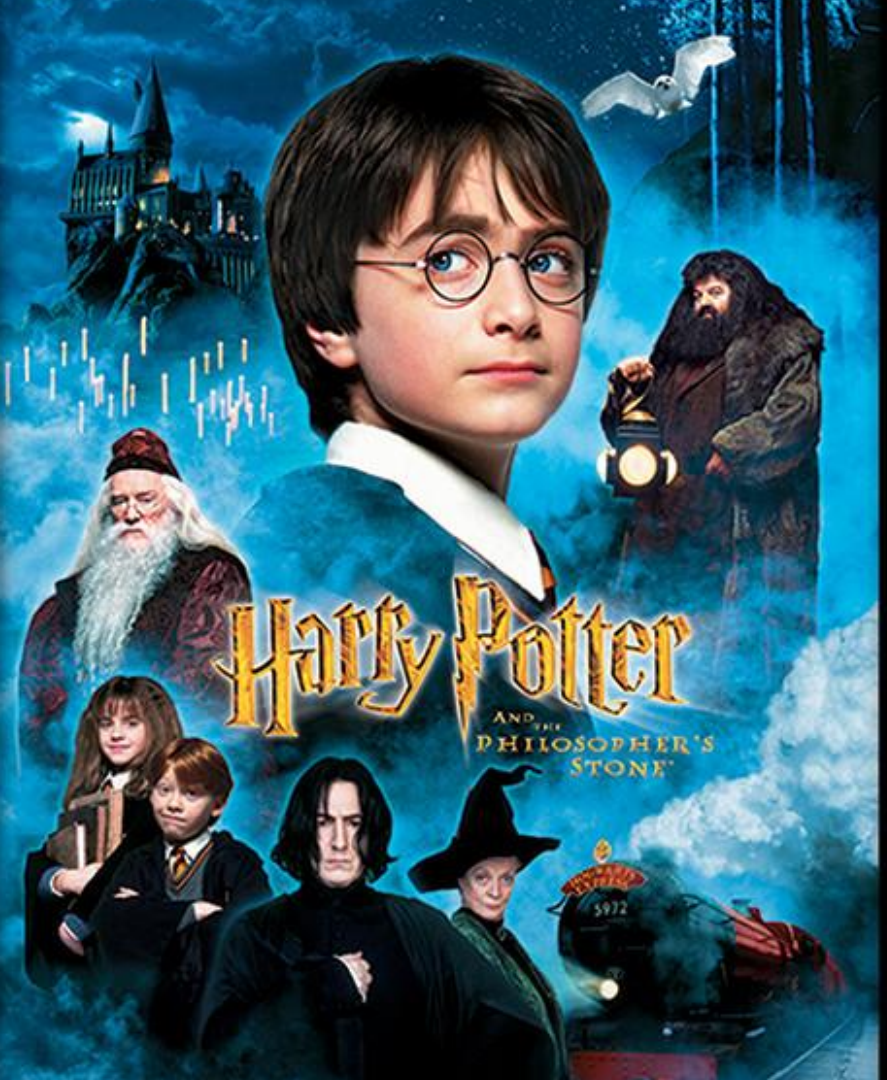
RANKING RECENT MOVIES

MOST PROMINENT DIRECTOR

CHRISTOPHER NOLAN

MOST FAMOUS ACTOR

JOHNNY DEPP

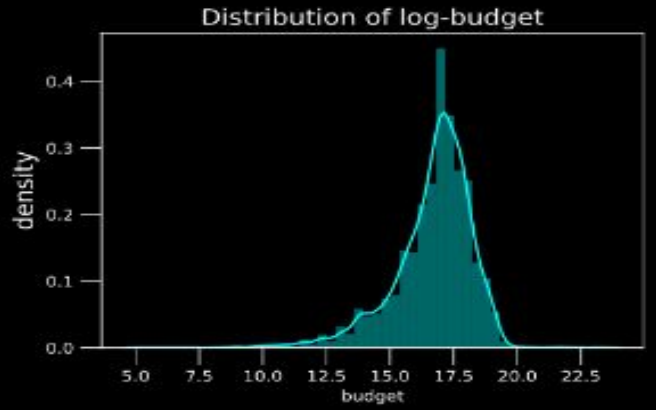


EXPLORATORY DATA ANALYSIS

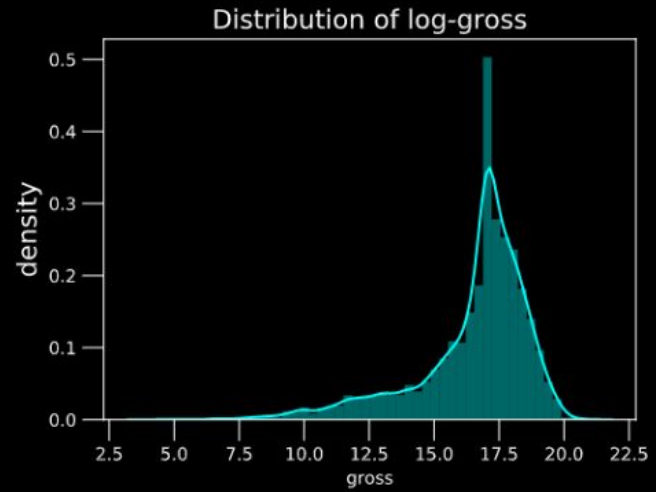
04

UNIVARIATE ANALYSIS

BUDGET

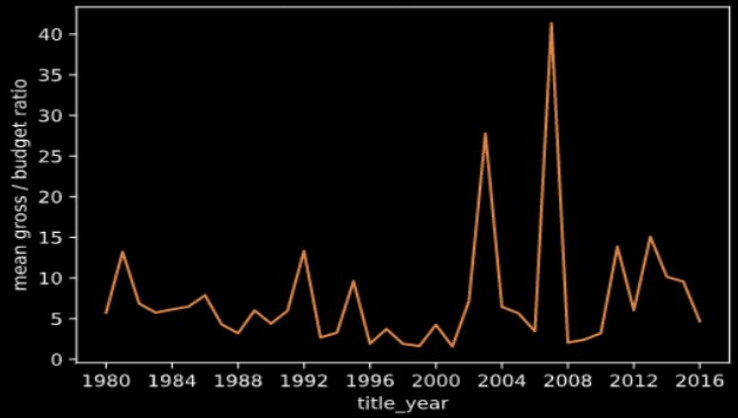


GROSS EARNINGS

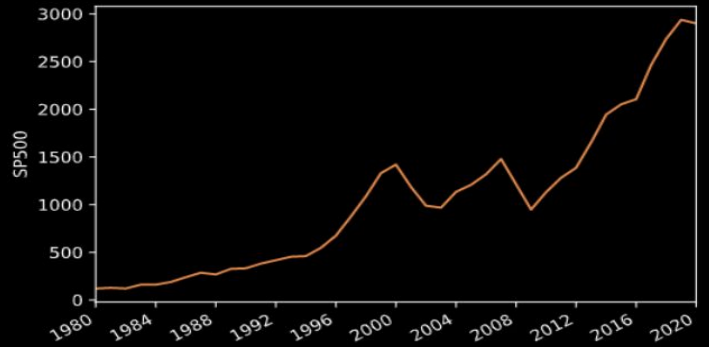


TREND WITH MACROECONOMIC VARIABLES

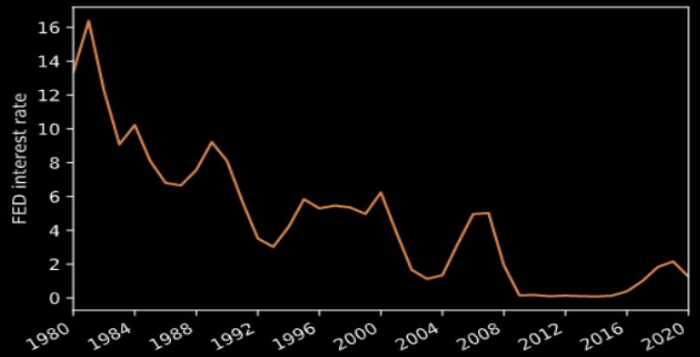
GROSS BUDGET RATIO



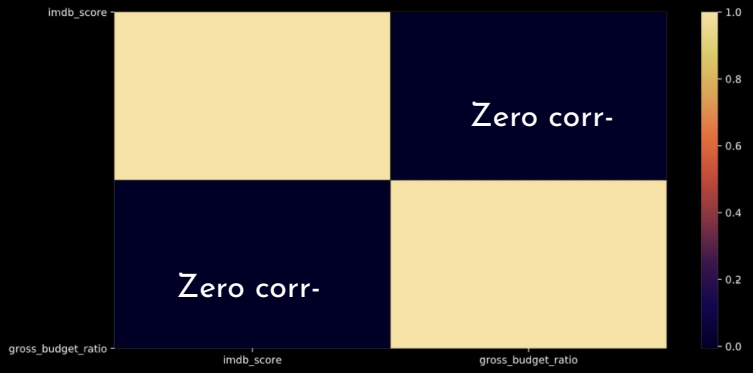
SP 500



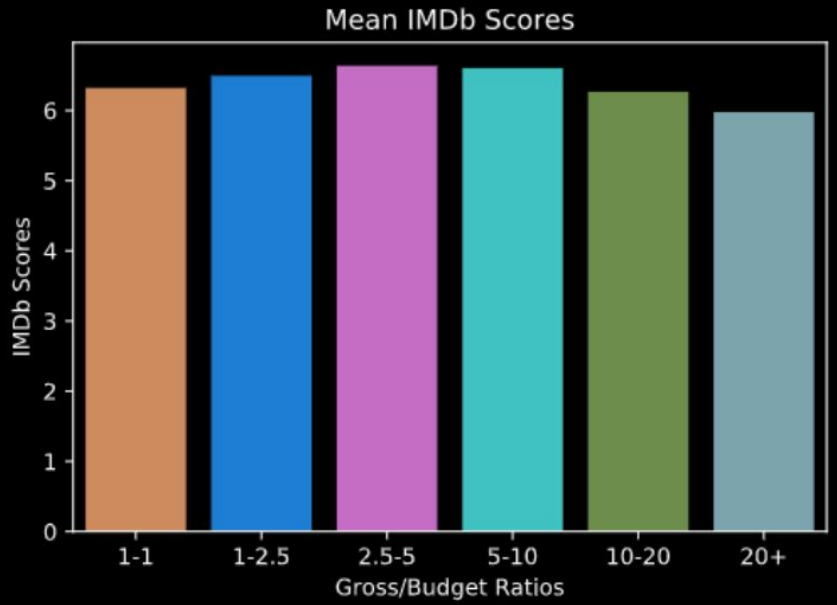
FED INTEREST RATE



BIVARIATE ANALYSIS



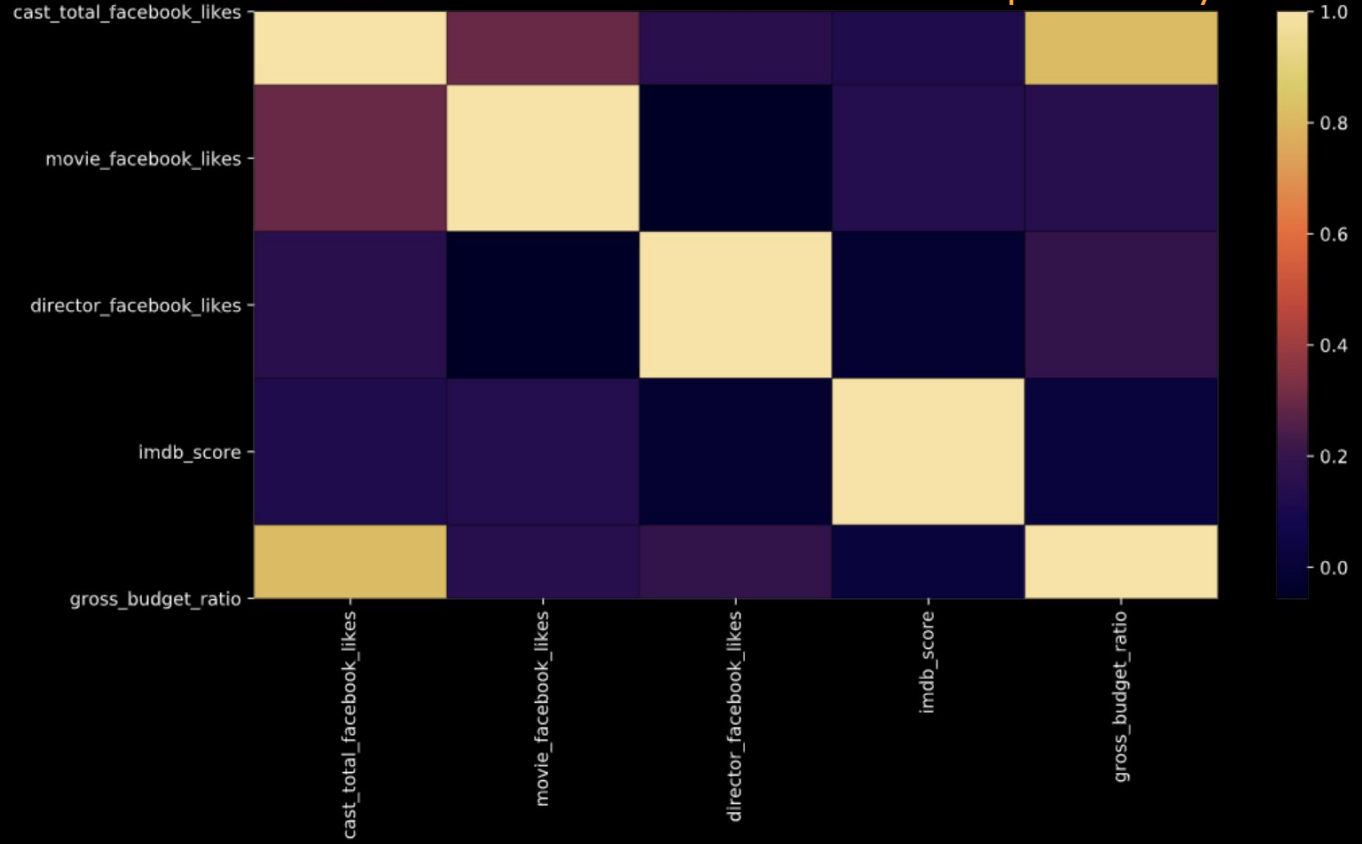
Correlation IMDb rating and gross/budget ratio



Distribution IMDb rating for groups of budget ratio

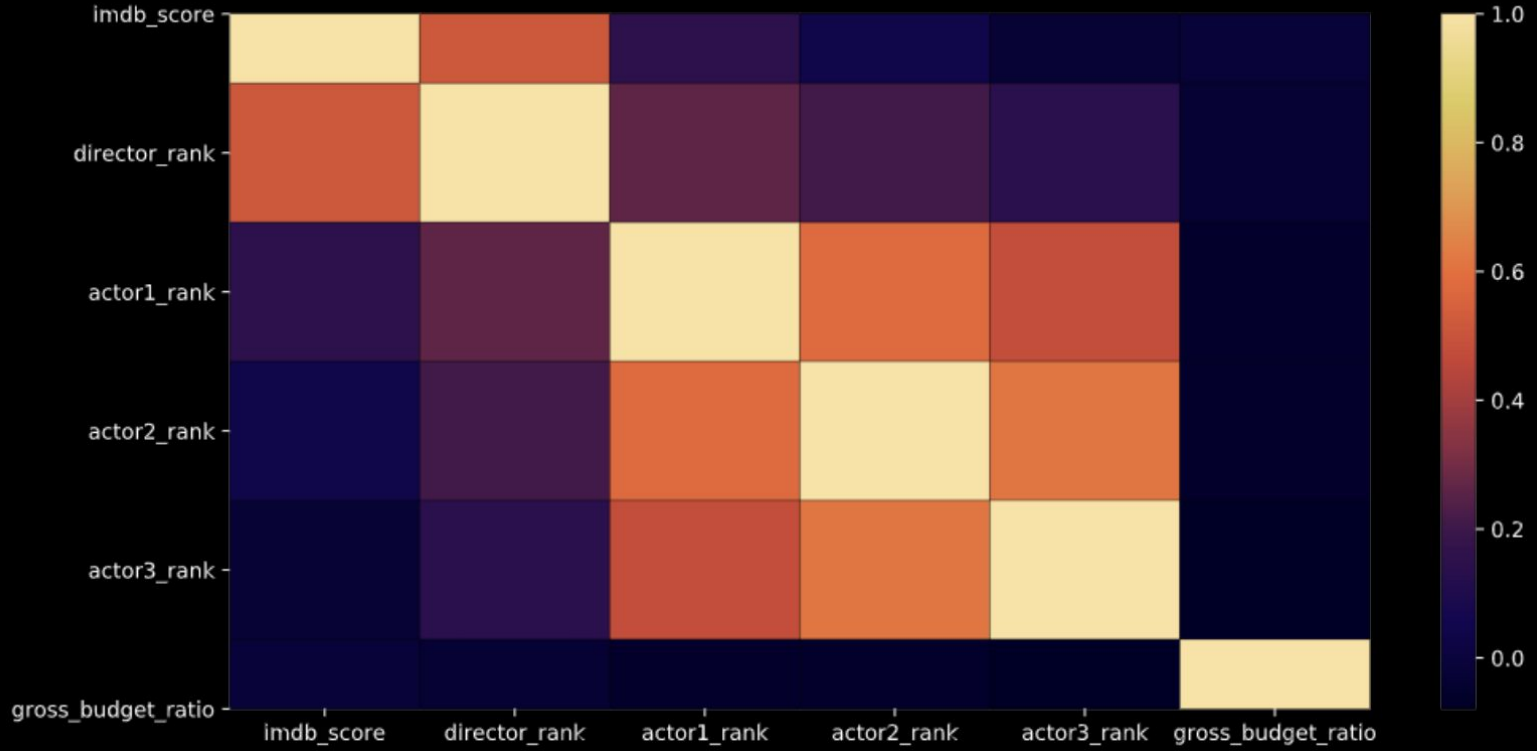
MULTIVARIATE ANALYSIS

Influence of Social Media on profitability

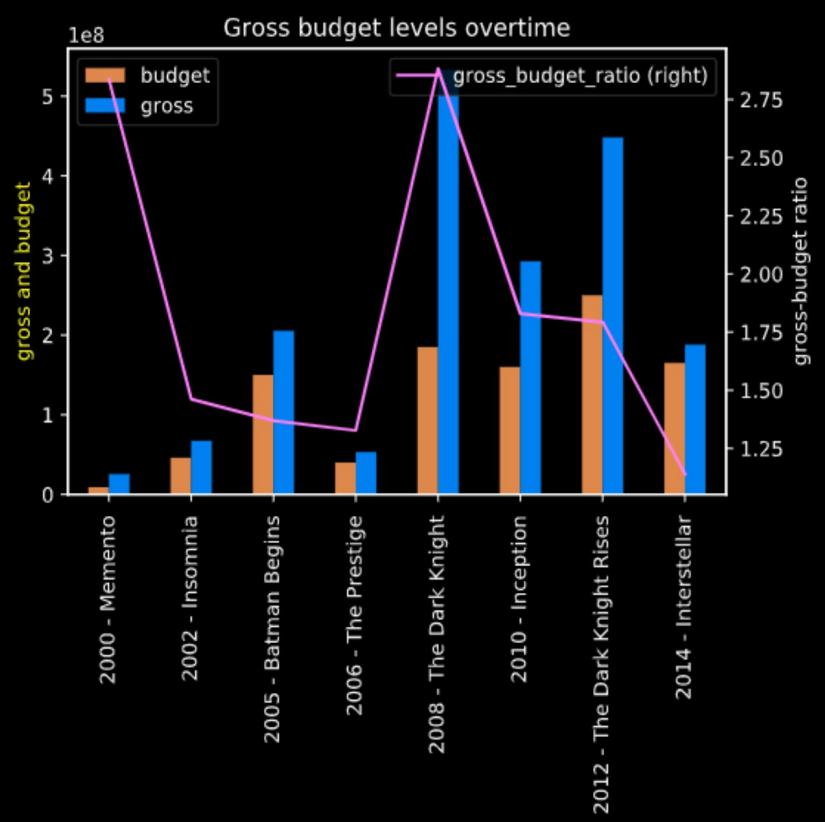


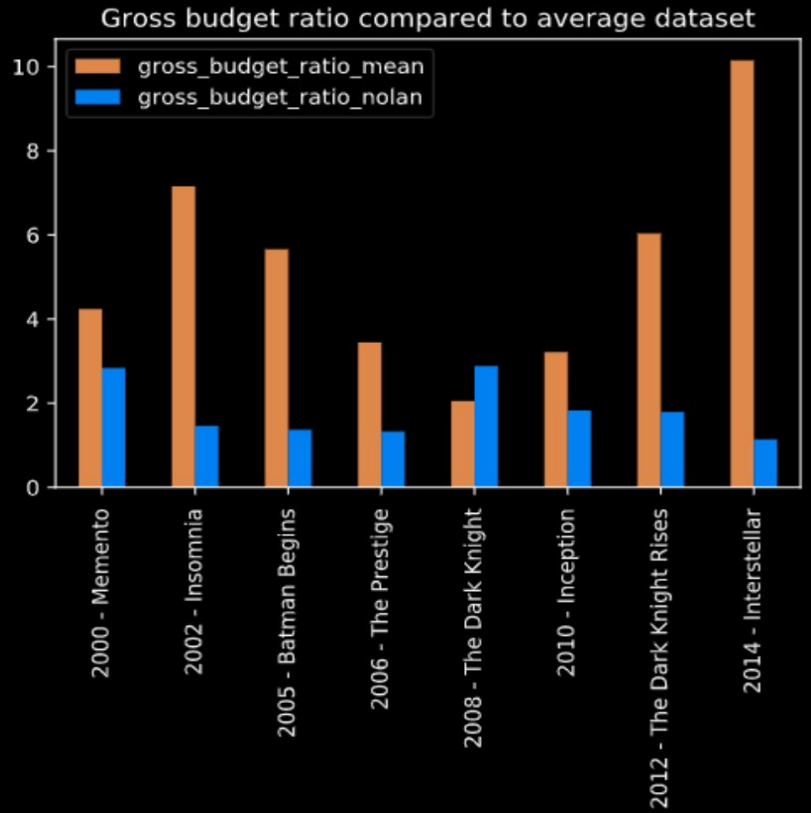
MULTIVARIATE ANALYSIS

Influence of cast and director on profitability and IMDb score



ANALYSIS PROMINENT DIRECTOR: CHRISTOPHER NOLAN

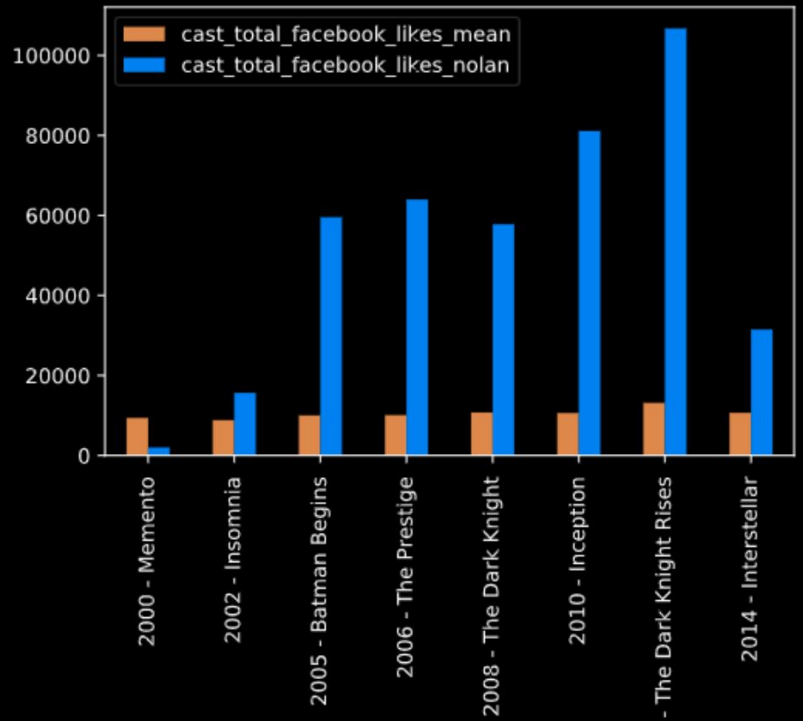




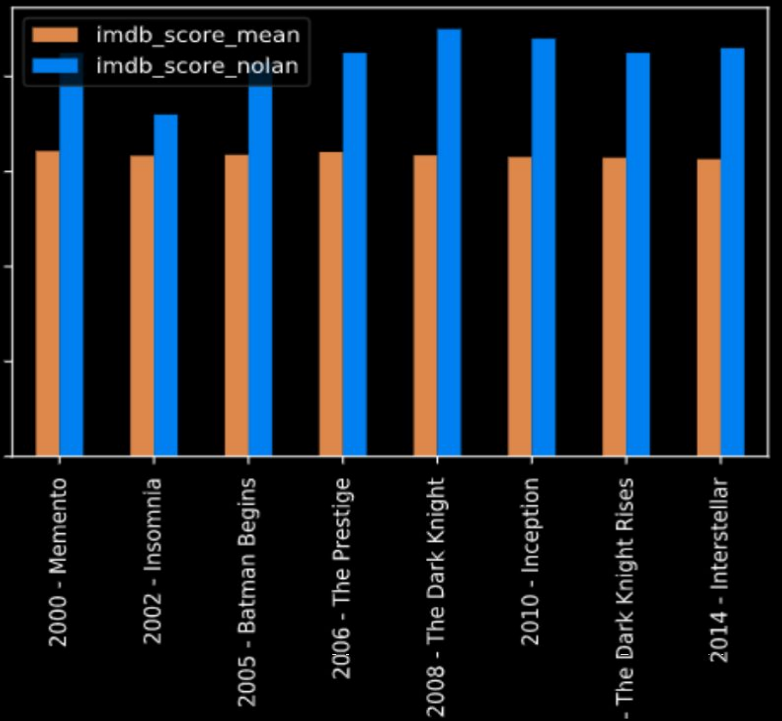
ANALYSIS PROMINENT DIRECTOR: CHRISTOPHER NOLAN

Difference between global average and Nolan's movies

Cast Facebook likes



IMDb score



PREDICTION MODELS & RESULTS

05

- For IMDb score we choose features available before and after release of the film including genre, rankings and social media response.
- In total 20 features can be used in prediction.

Duration	Budget	Director rank	Actor 1 rank	Actor 2 rank
Actor 3 rank	Movie FB likes	Cast Total FB likes	Biography	Comedy
Crime	Drama	Romance	Mystery - Thriller - Horror	Sci-Fi - Fantasy
Family - Animation	Action - Adventure	History - War	Others	Number of Faces on poster

FEATURE SELECTIONS

FEATURE SELECTIONS

- For the gross-budget ratio prediction, features available before the release of movies are selected.
- In total 26 features can be used in prediction.

Duration	Actor 1 FB Likes	Actor 2 FB Likes	Actor 3 FB Likes	Director FB Likes
Language	Cast Total FB Likes	Face Number in Poster	Country	Content Rating
Budget	Biography	Comedy	Crime	Drama
History - War	Romance	Mystery - Thriller - Horror	Sci-Fi - Fantasy	Family - Animation
Action - Adventure	Others	Actor 1 Rank	Actor 2 Rank	Actor 3 Rank
		Director Rank		

ADDITIONAL OPERATIONS

SCALING THE DATA

- Scaling the data for scale variant models such as KNN.
- Min-Max Scaler used.

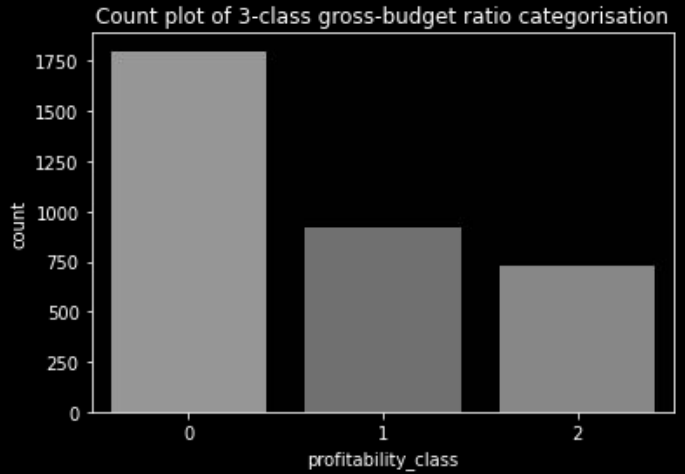
LOG OF GROSS-BUDGET RATIO

- Applied for some models.
- Due to highly skewness of the gross-budget ratio

HYPER - PARAMETER OPTIMIZATION

- Grid Search CV
- K-Fold Cross Validation

- For gross-budget ratio, we **eliminated instances with a ratio > 5** (too extreme outliers).
- Also for gross-budget ratio, we did predictions **through both regression and classification**, due to initial poor performance of the regression models.
- In classification of gross-budget ratio, first the classes were 0 (not profitable) and 1 (profitable); then it was 0 (not profitable), 1 (profitable), and 2 (very profitable).
- Three-class classification suffered from the problem of **data imbalance**.



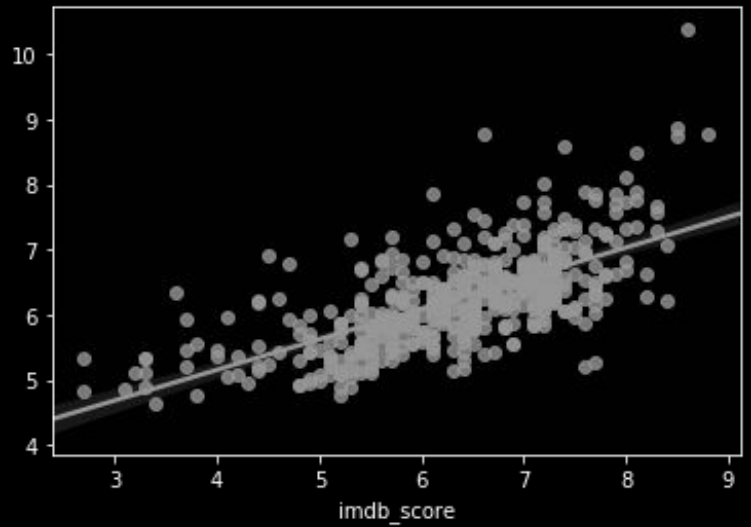
ADDITIONAL OPERATIONS

- We tried to predict with the following algorithms:

Algo	Variations	Regression (for IMDB score AND gross-budget ratio)	Classification (for gross-budget ratio)
Linear Regression	OLS, Ridge, Lasso, ElasticNet	Yes	No
Logistic Regression	Binary and Multi-Class	No	Yes
Decision Trees	Ordinary DT, Random Forest, XGBoost	Yes	Yes
K-nearest Neighbour	KNN	No	Yes
Neural Networks	Semu, Sigmoid	Yes	Yes

LINEAR REGRESSION (OLS)

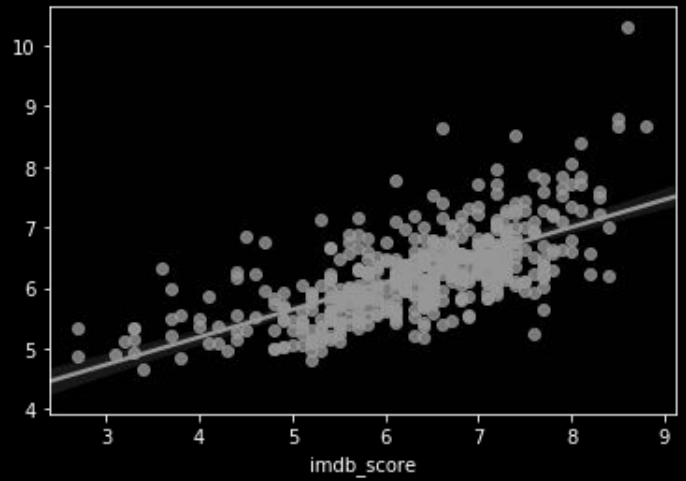
- Used for: IMDb score.
- Finds the unbiased coefficients that best fit the data given.
- Advantages: no parameters, small execution time, simplicity of changes detection.
- Metrics: MSE, R2.
- Best performance results:



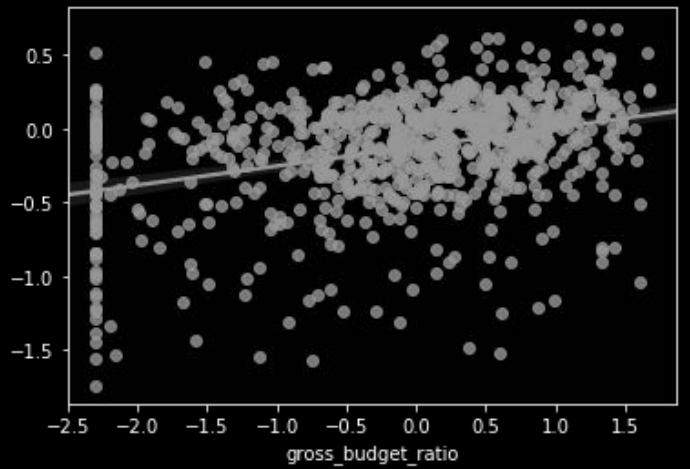
MSE	R2
0.64	0.44

LINEAR REGRESSION (RIDGE)

- Used for: IMDb score, gross-budget ratio.
- Regularized linear regression method, which works better if features have multicollinearity.
- Advantages: decreases the less significant features coefficients, reduces variation
- Metrics: MSE, R2 -------- Parameter: alpha
- Best performance results:



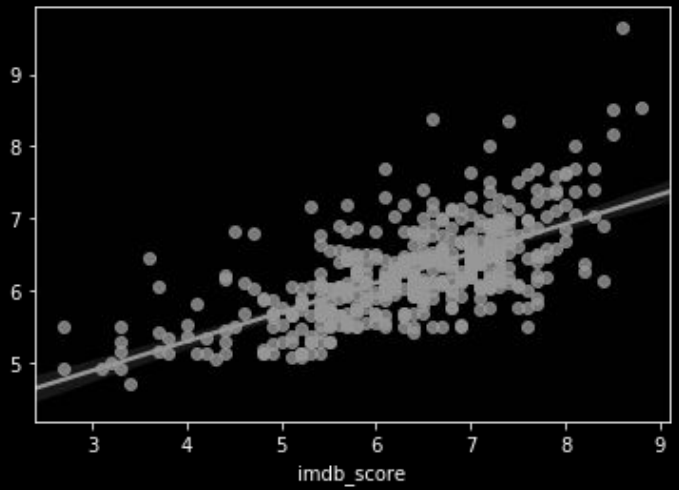
MSE	R2
0.63	0.45



MSE
0.96

LINEAR REGRESSION (LASSO)

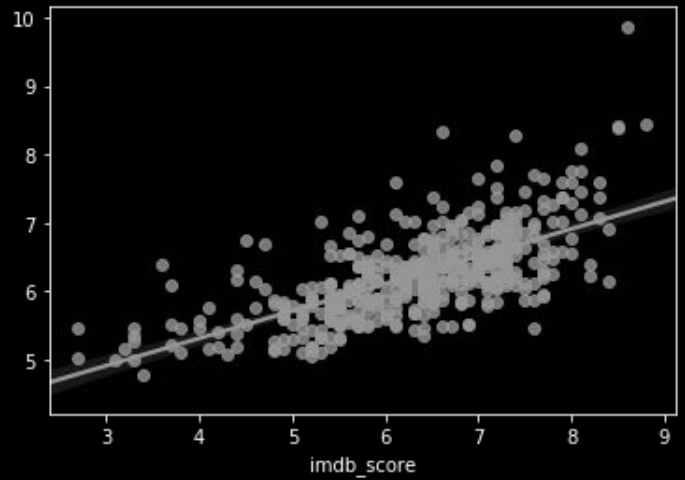
- Used for: IMDb score.
- Regularized linear regression algorithm, which is similar to Ridge regression, making non-significant coefficients equal 0.
- Advantages: creates models with fewer features (scarce models)
- Metrics: MSE, R2. -------- Parameter: alpha
- Best performance results:



MSE	R2
0.65	0.43

LINEAR REGRESSION (ELASTIC NET)

- Used for: IMDb score.
- Regularized linear regression model, combines feature elimination property of Lasso and feature coefficient reduction property of from the Ridge model.
- Advantages: creates models with fewer features, reduces complexity
- Metrics: MSE, R2. -------- Parameter: alpha
- Best performance results:



MSE	R2
0.65	0.44

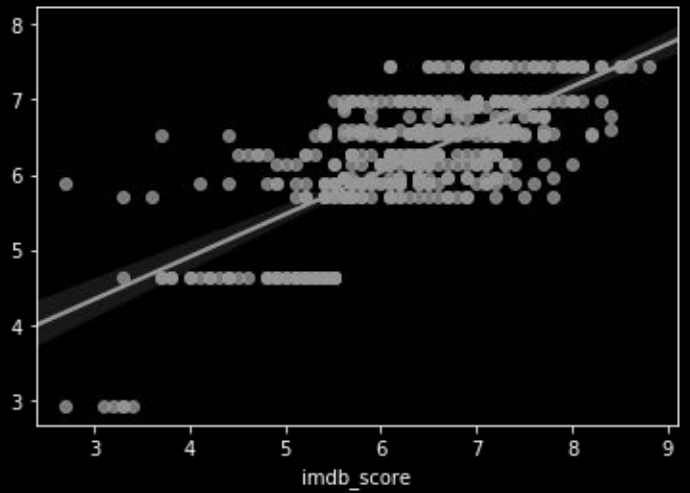
- Used for: gross-budget ratio classification.
- Supervised learning method, can be used for binary and multi-class classification. Optimizes hyper-parameter 'C' (higher regularization).
- Higher C - less regularization, lower C - more regularization
- Metrics: accuracy score, specificity score, recall score.
- Best performance results:

specificity score	accuracy score	recall score
0.55	0.62	0.69

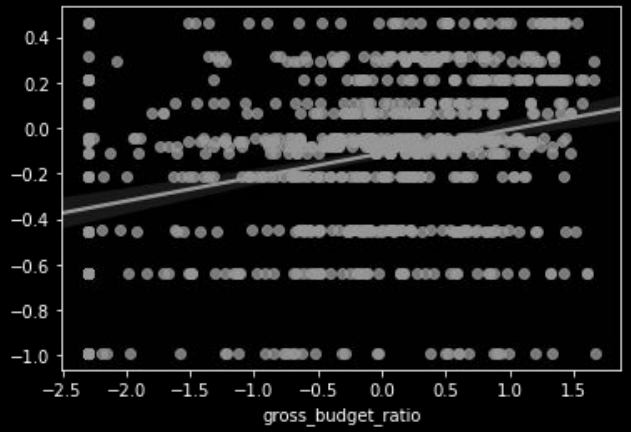
LOGISTIC REGRESSION

DECISION TREES (ORDINARY)

- Used for: IMDb score and gross-budget ratio.
- It tries to solve the problem by using a tree representation. In the tree, each internal node of the tree corresponds to a feature condition, and each leaf node corresponds to a class label (classification) or a value (regression)
- Metrics: MSE, R2 (Regression)- accuracy score, specificity score, recall score. (classification)
- Best performance results:



MSE	R2
0.57	0.51



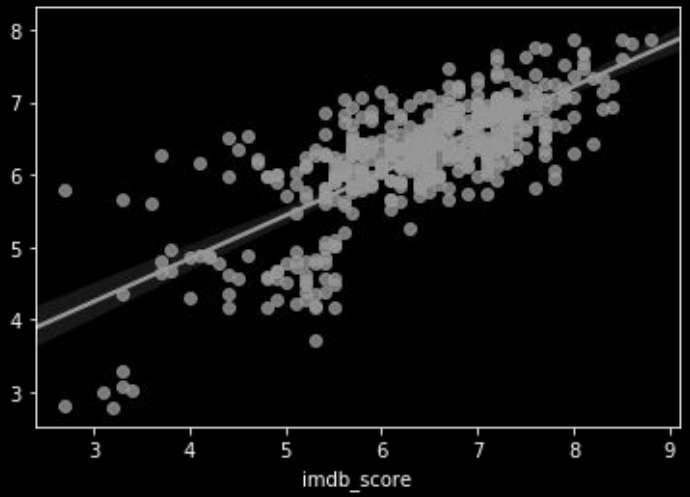
MSE
0.93

specificity score	accuracy score	recall score
0.68	0.57	0.47

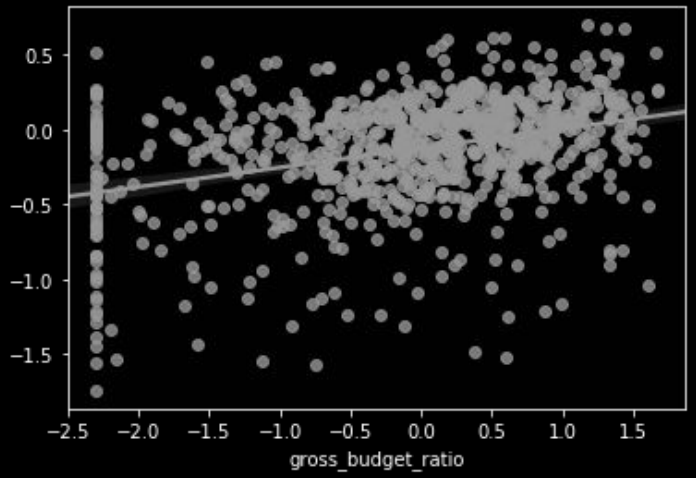
DECISION TREES (ORDINARY)

RANDOM FOREST

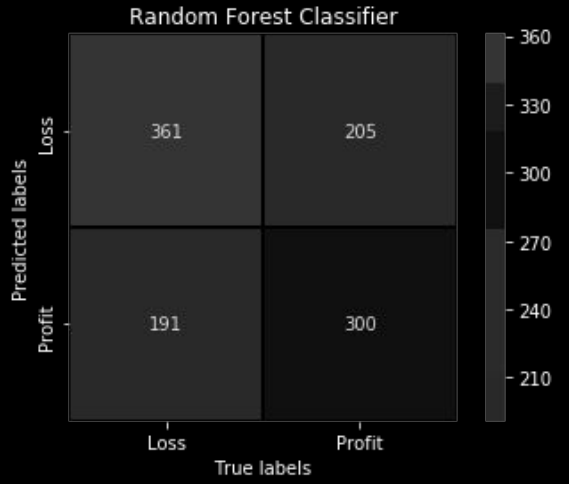
- Used for: IMDb score and gross-budget ratio.
- Compared to the decision tree algorithm using full data set, Random Forest randomly selects observations and features and creates many decision trees then takes average
- Metrics: MSE, R2 (Regression)- accuracy score, specificity score, recall score. (classification)
- Best performance results:



MSE	R2
0.50	0.56



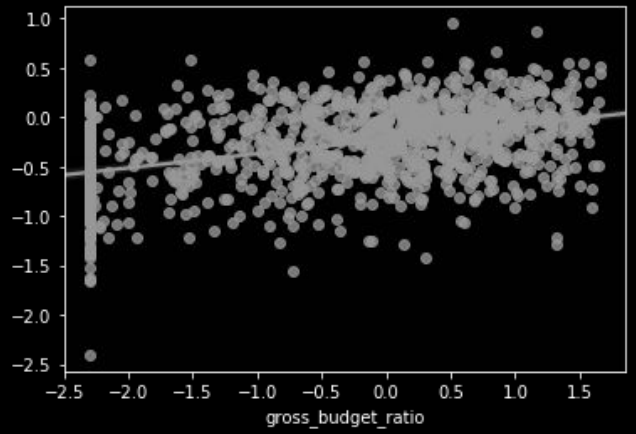
MSE
0.90

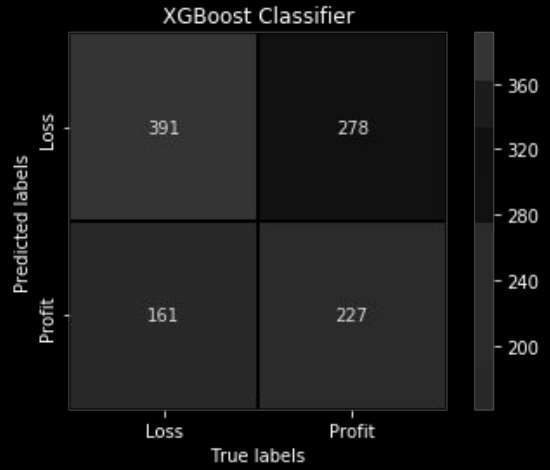


specificity score	accuracy score	recall score
0.65	0.63	0.59

RANDOM FOREST

- Used for: Only for best performing model of gross-budget ratio prediction.
- Long computational time
- Metrics: MSE, R2 (Regression)- accuracy score, specificity score, recall score. (classification)
- Best performance results:



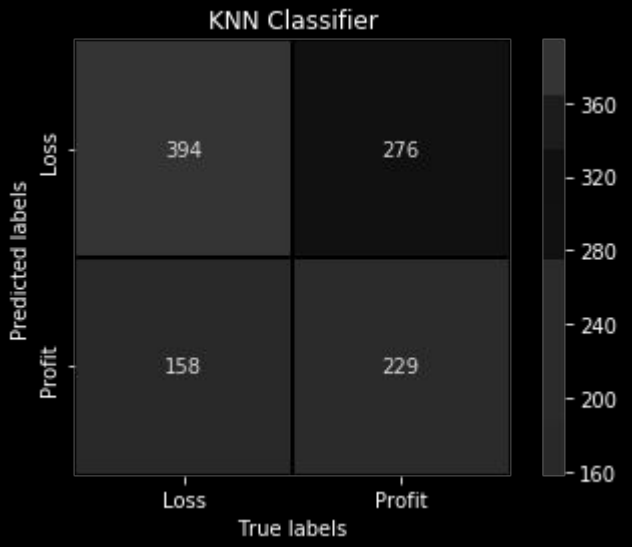


XGBOOST

MSE
1.09

specificity score	accuracy score	recall score
0.71	0.58	0.45

- Used for: Gross-Budget ratio classification
- Predicts the output based on the similarity function which is calculated with k nearest observations.
- Lower k - risk of overfit, high k - chance of underfit
- Metrics: Accuracy, Specificity and Recall Score
- Best performance results:



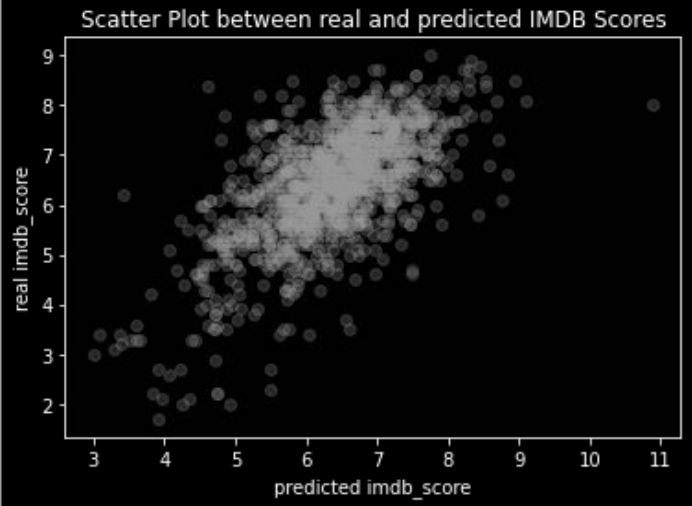
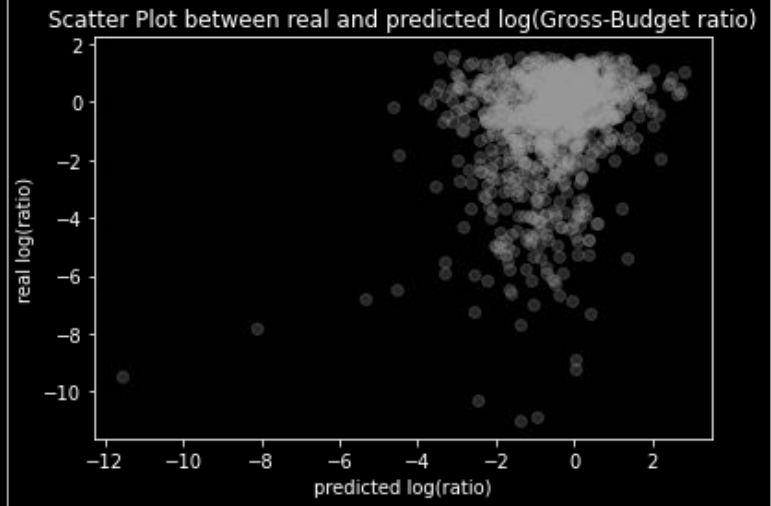
specificity score	accuracy score	recall score
0.71	0.59	0.45

K-MEANS CLUSTERING

Network Architecture	1. Regressing IMDB Score	2. Regressing log(Gross-budget Ratio)	3. Classifying Gross-budget Ratio
# hidden dense layers	1	1	1
# neurons in hidden layer(s)	32	32	32
Activation function	selu	selu	sigmoid, softmax
Optimiser	Adam	Adam	Adam
Loss function	MSE	MSE	Cross entropy
Miscellaneous		log (ratio) was taken due to the heavily skewed data.	Used drop out for the binary clf to avoid overfitting.

NEURAL NETWORKS

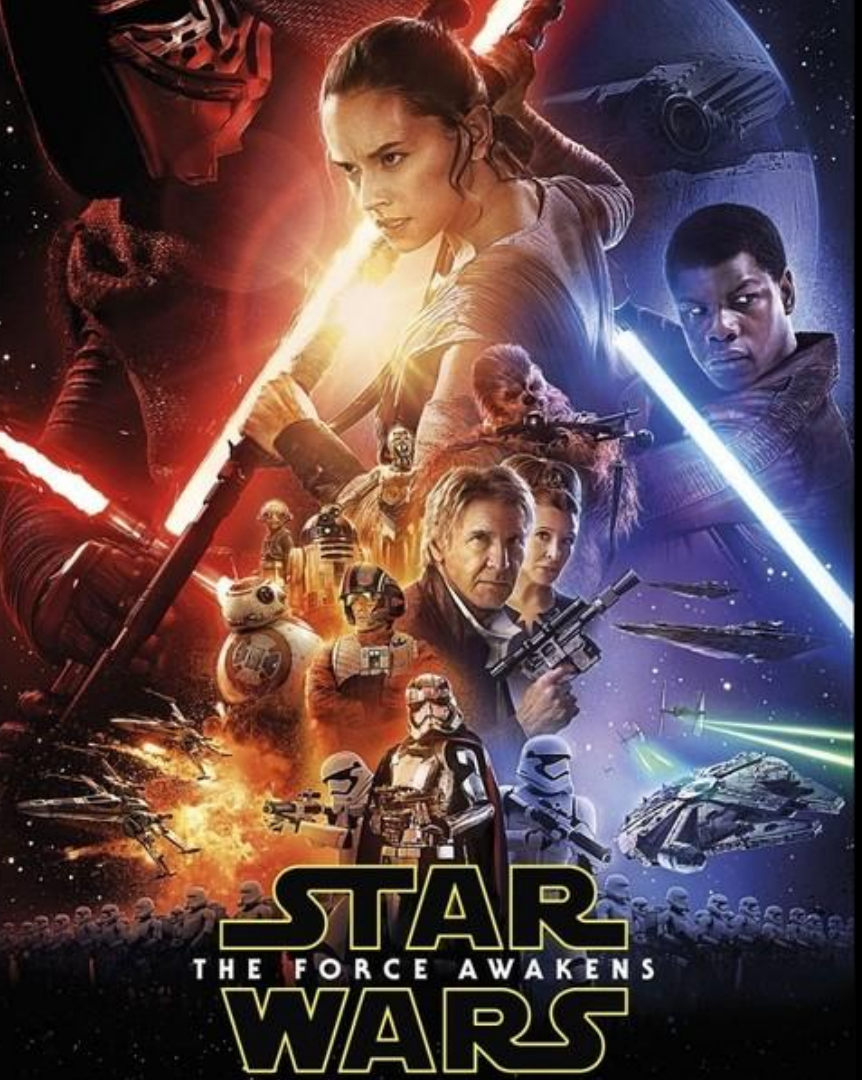
Results of Regression

	IMDB Score	log(Gross-budget Ratio)
MAE	0.670	1.361
MSE	0.767	3.541
		

NEURAL NETWORKS

Results of Classification

	Binary clf	3-class clf with resampling	3-class clf without resampling
Accuracy	0.6	0.50	0.53
Precision	0.63 / 0.56	0.59 / 0.35 / 0.43	0.76 / 0.24 / 0.22
Recall	0.64 / 0.55	0.67 / 0.33 / 0.35	0.61 / 0.29 / 0.42
F1-score	0.64 / 0.56	0.62 / 0.34 / 0.39	0.67 / 0.26 / 0.29

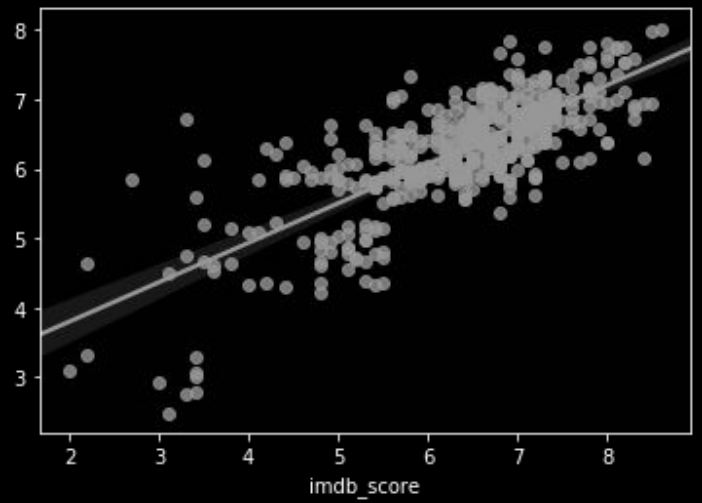


INTERPRETATION OF THE RESULTS

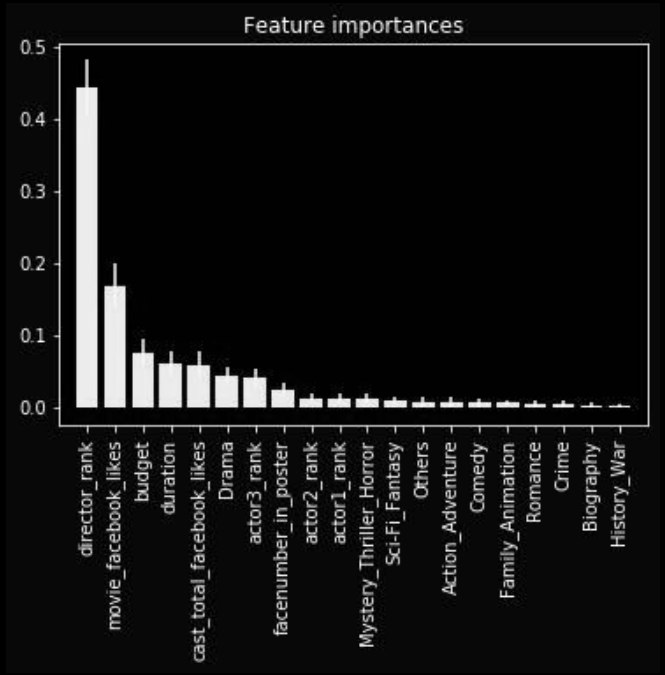
06

INTERPRETATIONS IMDb SCORE

- The best algorithm: Random Forest
- The best performing model: Films from 2009
- The most important features for Random Forest: director rank and movie FB likes

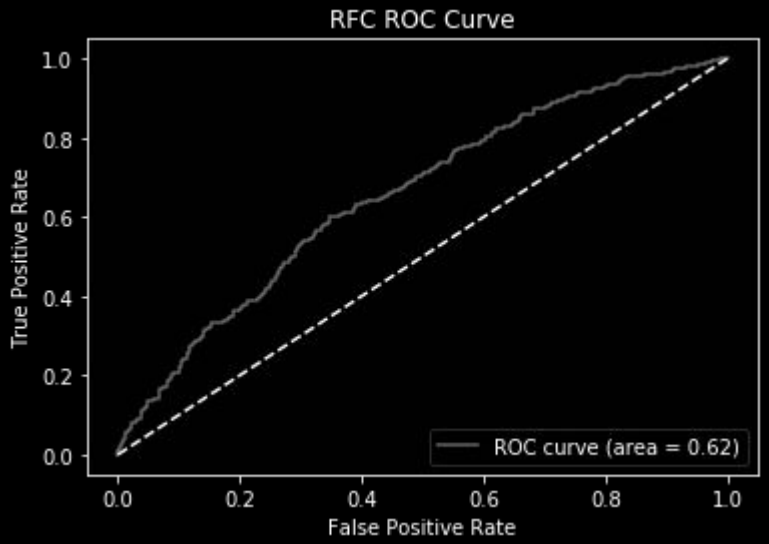


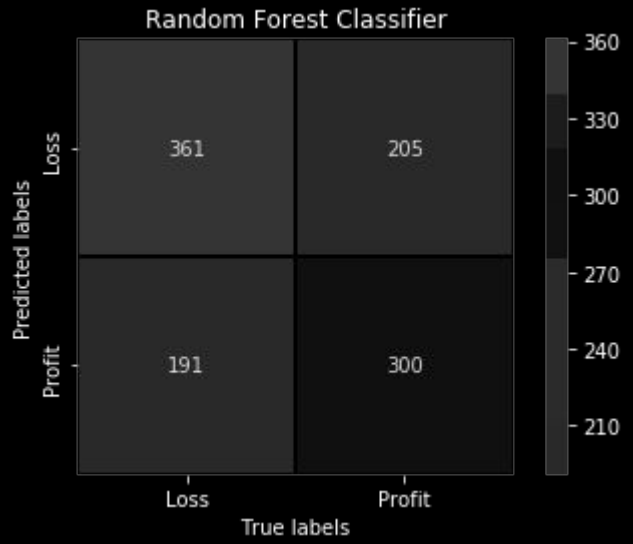
MSE	R2
0.5	0.56



INTERPRETATIONS GROSS-BUDGET CLASSIFICATION

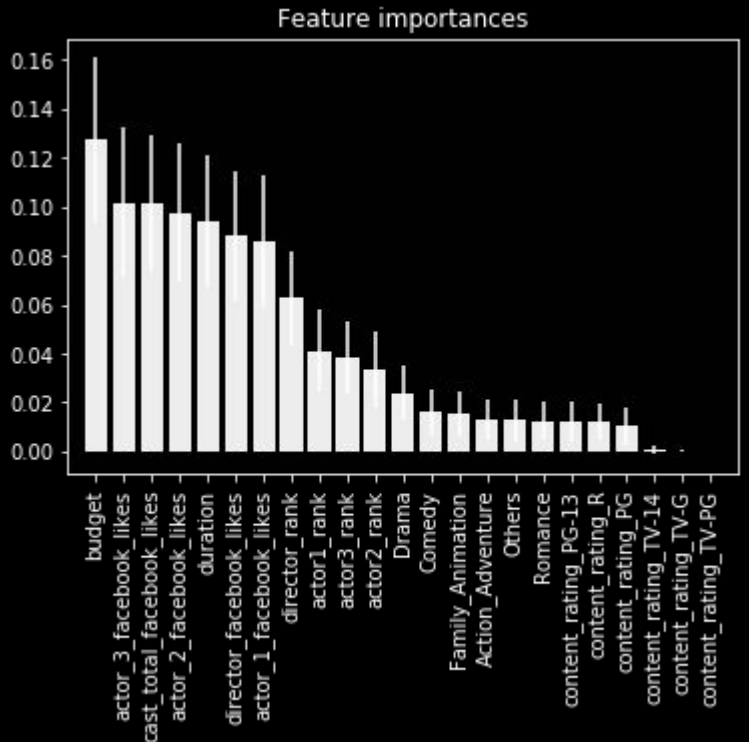
- The best algorithm: Random Forest
- The best performing model: Binary Classification on all movies with log transformation and feature elimination.
- Optimized hyper-parameters : max_depth = 10, n_estimators = 200
- Can be used by risk-averse movie producers.





INTERPRETATIONS GROSS-BUDGET RATIO

- As can be predicted, the most important feature is Budget.
- Actor 3 FB likes => surprising.



IMPLICATIONS FOR BUSINESSES

07

FEATURES RELEVANCE

- In the first part of the project, we tried to identify the most relevant features and their correlations. Surely, the two attributes at which the company should pay attention are
  - The **gross budget ratio**
  - and the **IMDb score**.
- But even if we try to assign a *rank* to the cast and the director, these features are not well correlated with profitability.

  However, even if the global ranking does not directly impact the success of the movie, when we analysed the **high-ranked director and actor**, we found their movies to be more successful than average; so considering also these feature for the predictions makes sense!

- Regarding *Social Media*, we demonstrated that there exists a high correlation between

  Gross Budget Ratio ⟷ Cast Total Facebook Likes

- *Budget* is obviously well correlated with the profitability by construction.
- => Movie businesses should pay particular attention to the Cast Facebook Likes and the Budget, as they do bear significant correlations to a film's commercial success.
- Considering that IMDb rating and profitability index are not at all correlated, we predicted them separately.

- In terms of prediction models, we recommend businesses to make use of the Random Forest algorithm, as this model is not only intuitive but it has proven to also be robust in its performance.

- For the gross-budget ratio prediction, movie producers can use the best performing binary classification method if they are risk-averse. Otherwise, if they are risk loving, we wouldn't suggest using it because it predicts 40% of profitable movies as loss.

- However, the poor performance observed in predicting profitability has shown that this variable may be unpredictable or more features/more engineering need to come into play to make this task possible!

IMPLICATIONS

“

Audience can live without a movie but a movie cannot live without an audience..”

—Amit Kalantri