Predicting Movie Success

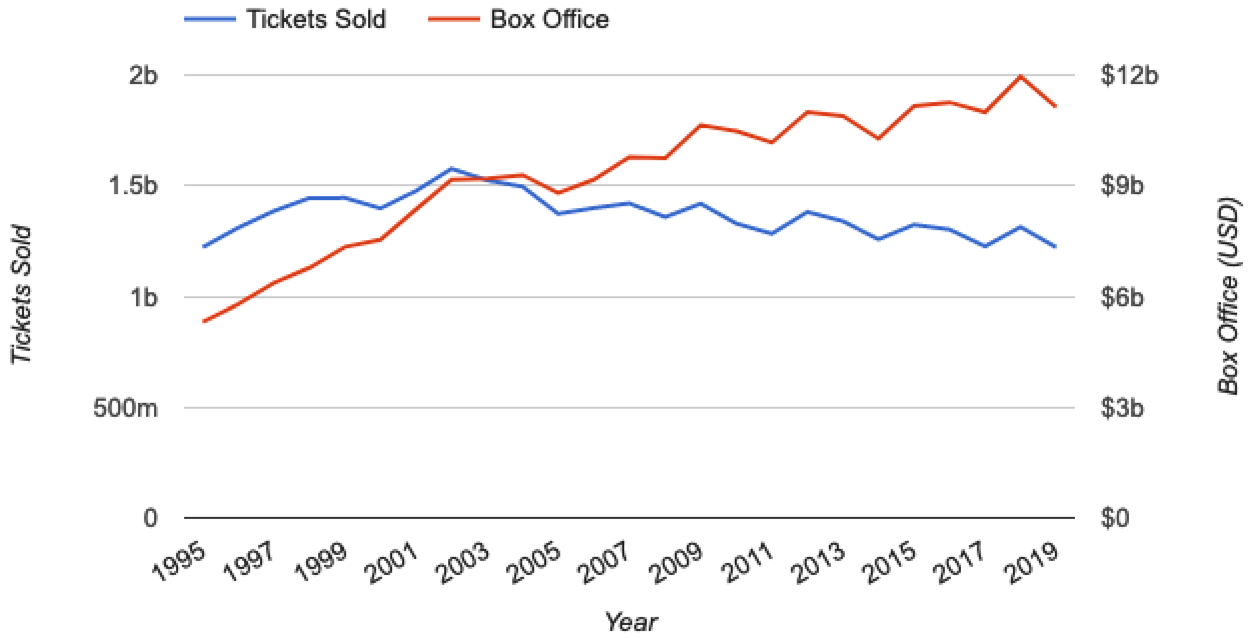
DATS6103 Project Presentation
Amna Gul, Hemanth Koganti , Madhuri Yadav

Agenda

- ❑ Introduction - Problem Statement
- ❑ Data
- ❑ Preprocessing
- ❑ EDA
- ❑ Modeling
- ❑ Conclusion
- ❑ References

❑ Introduction

- Film industry is one of the top grossing industries in the world



Problem Statement

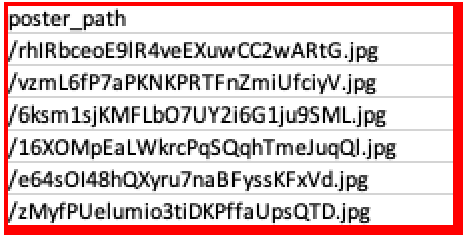
- Billions of dollars are invested each year, expecting high margin of profit
- Whether is it possible to use machine learning algorithms to predict if movie will be a success?
    - Criteria for success
        - Movie is able to generate amount of revenue that is “greater” than the budget of the film

Data Source

- Primary source of our data set is Kaggle (TMDB) till July 2017
- Columns were also added by parsing data made publicly available on IMDb’s website
    - Recent data 2019

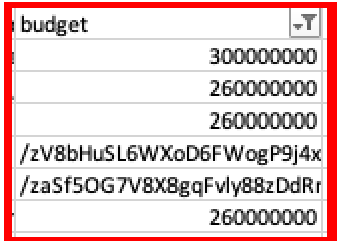
Data Preprocessing/Cleaning

- Raw data: 45,000 rows and 24 columns
  - ➢ Remove irrelevant columns e.g. “home page”, “poster path”



poster_path
/rhIRbceoE9lR4veEXuwCC2wARtG.jpg
/vzmL6fP7aPKNKPRTFnZmiUfciyV.jpg
/6ksm1sjKMFLbO7UY2i6G1ju9SML.jpg
/16XOMpEaLWkrcPqSQqhTmeJuqQl.jpg
/e64sOI48hQXyru7naBFyssKFxVd.jpg
/zMyfPUelumio3tiDKPffaUpsQTD.jpg

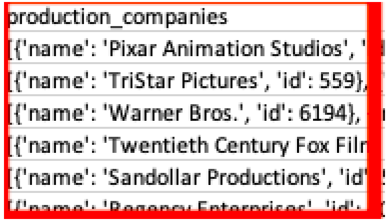
  - ➢ Searched for corrupt values in remaining columns e.g. “budget” containing alpha-numeric values



budget
300000000
260000000
260000000
/zV8bHuSL6WXoD6FWogP9j4x
/zaSf5OG7V8X8gqFvly88zDdRr
260000000

Data Preprocessing/Cleaning

- Columns “Genre” and “Production_Companies” were in JSON format



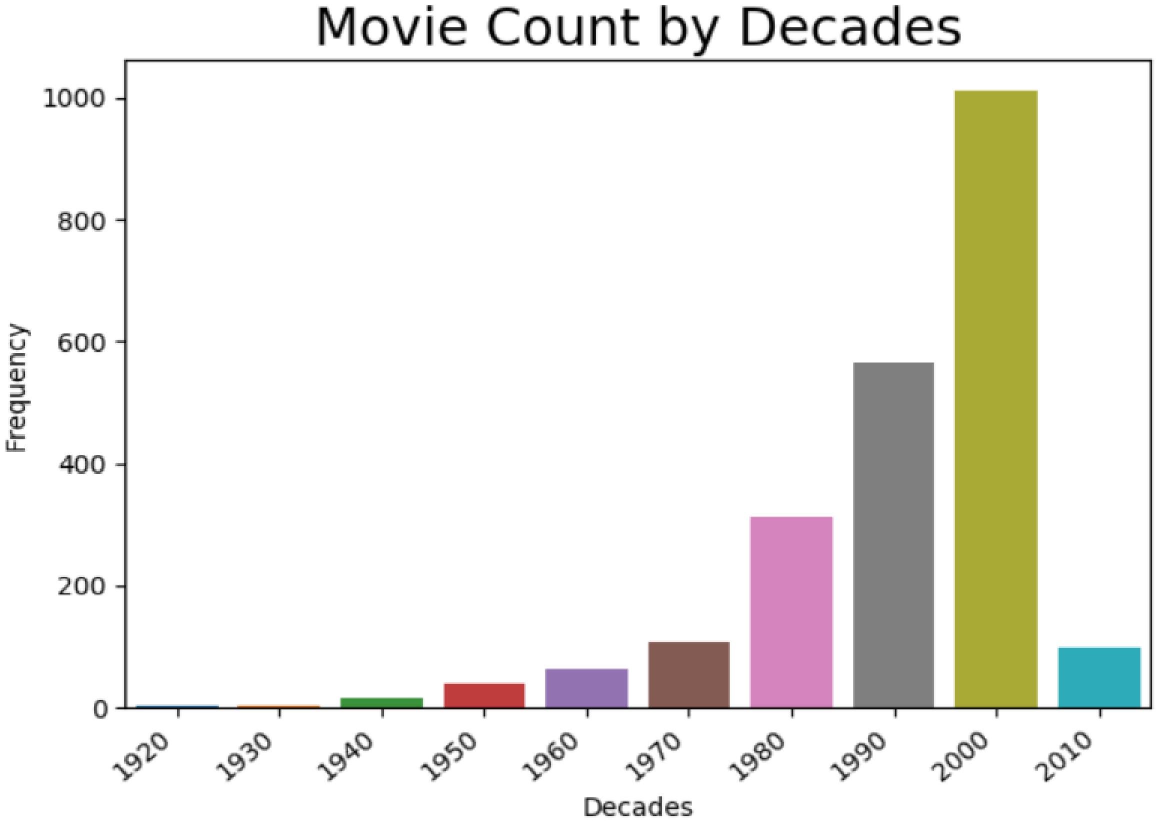
production_companies
[{'name': 'Pixar Animation Studios', '
[{'name': 'TriStar Pictures', 'id': 559},
[{'name': 'Warner Bros.', 'id': 6194},
[{'name': 'Twentieth Century Fox Filr
[{'name': 'Sandollar Productions', 'id'
[{'name': 'Regency Enterprises', 'id':

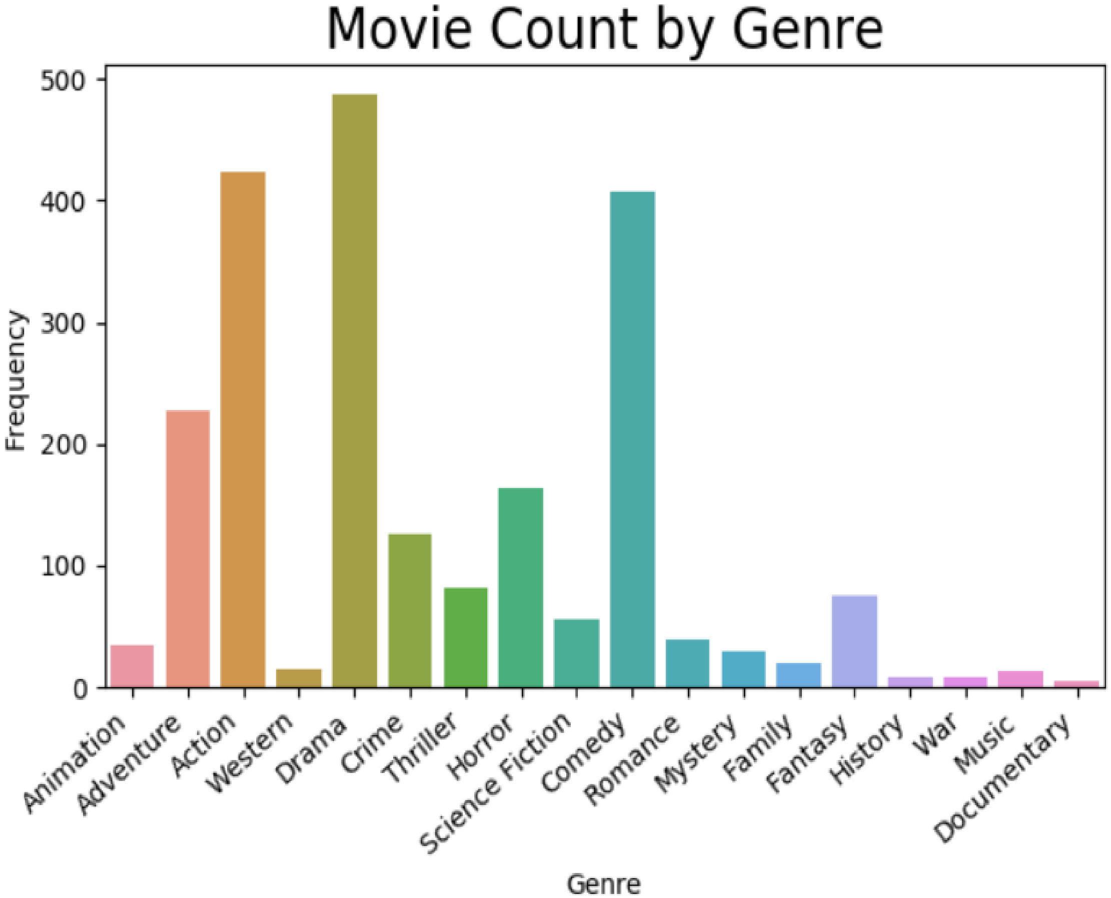
- Converted columns to their proper data type e.g. “release_date” was converted date-time format instead of string. Month was extracted to create a separate column.
- Merged “average_rating” and “vote_count” from IMDb’s website

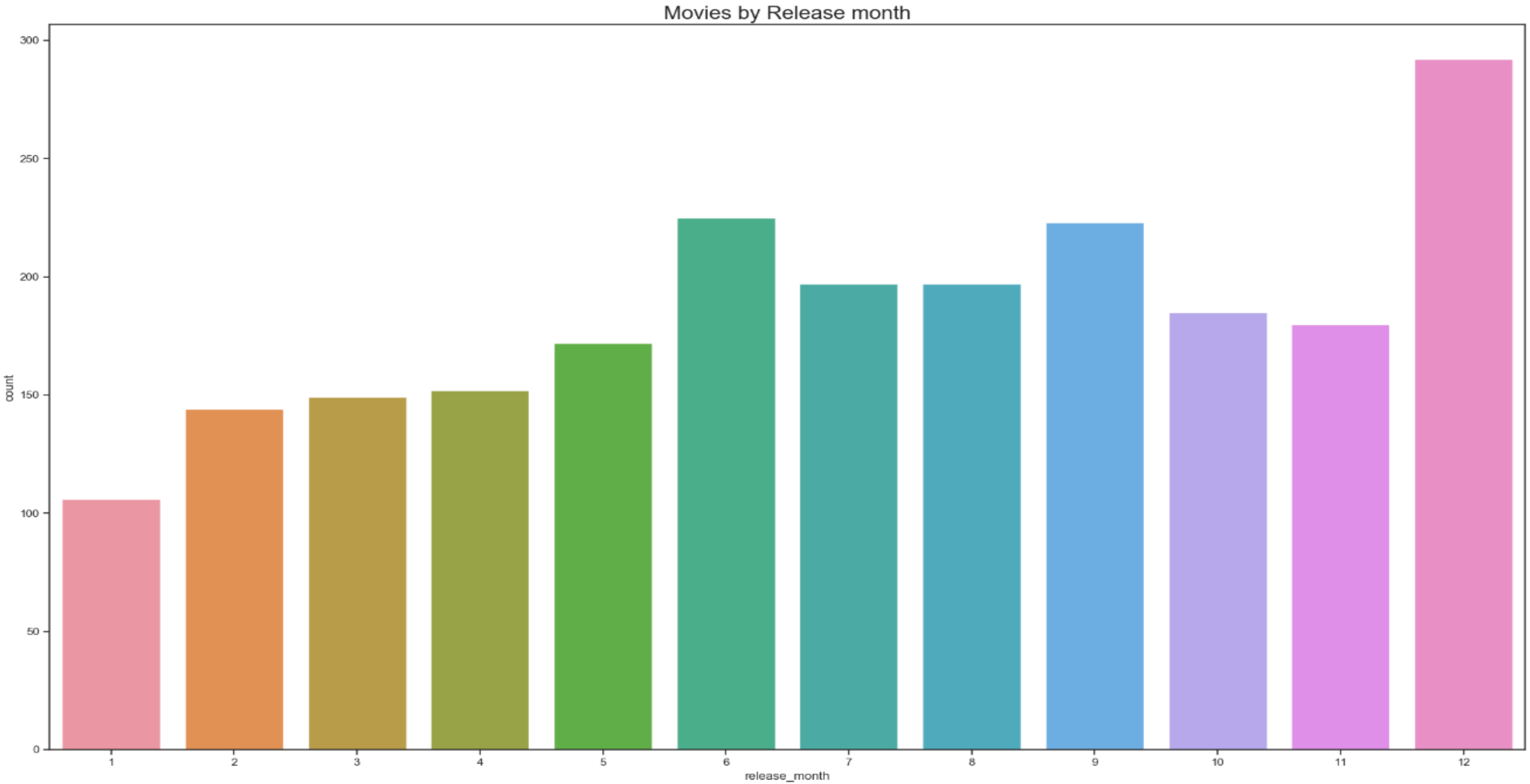
Data Preprocessing/Cleaning

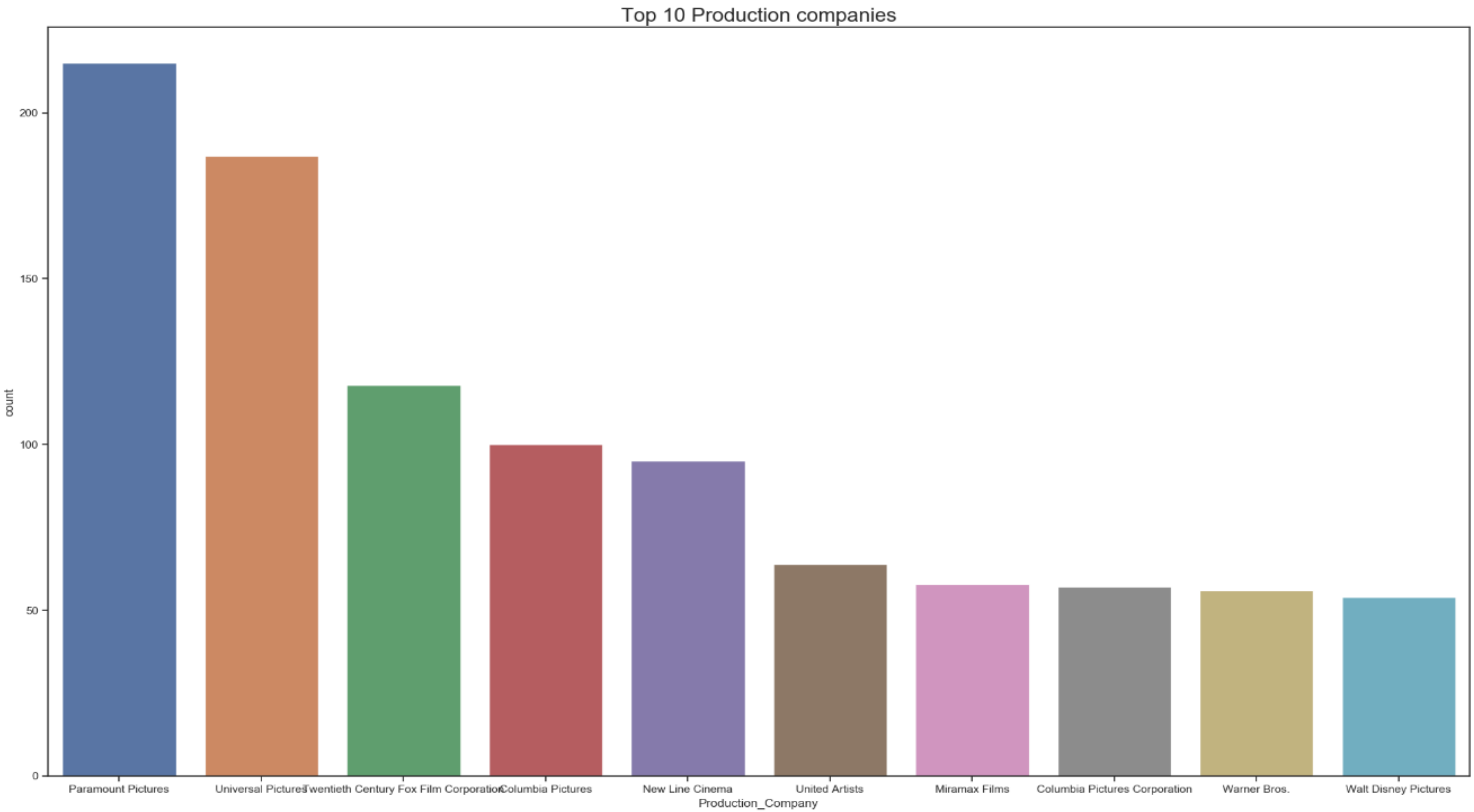
- ➢ Outliers: excluded all those rows for which budget or revenue value was unrealistically low
- ➢ Created our target column by dividing revenue by budget
- ➢ Removed duplicates
- ➢ End result: 2,222 rows & 17 columns

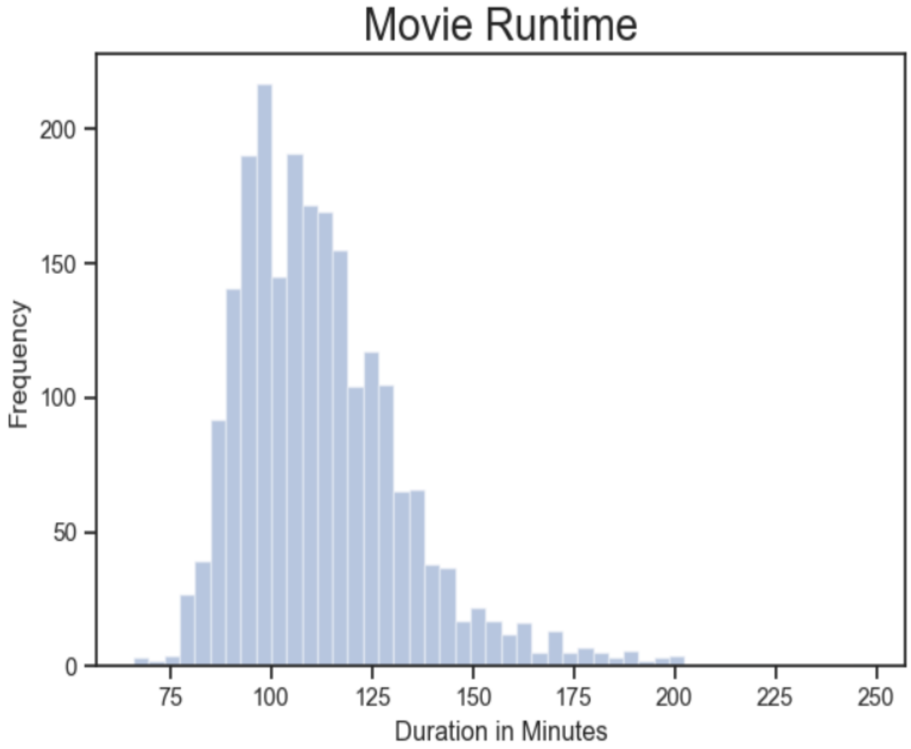
EDA

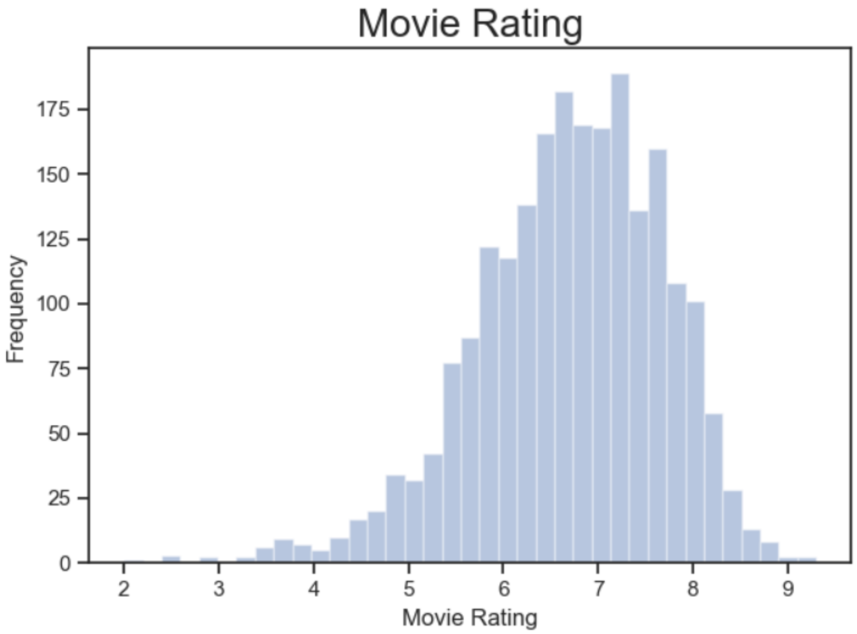

Movie Count by Decades
Frequency
1000
800
600
400
200
0
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
Decades

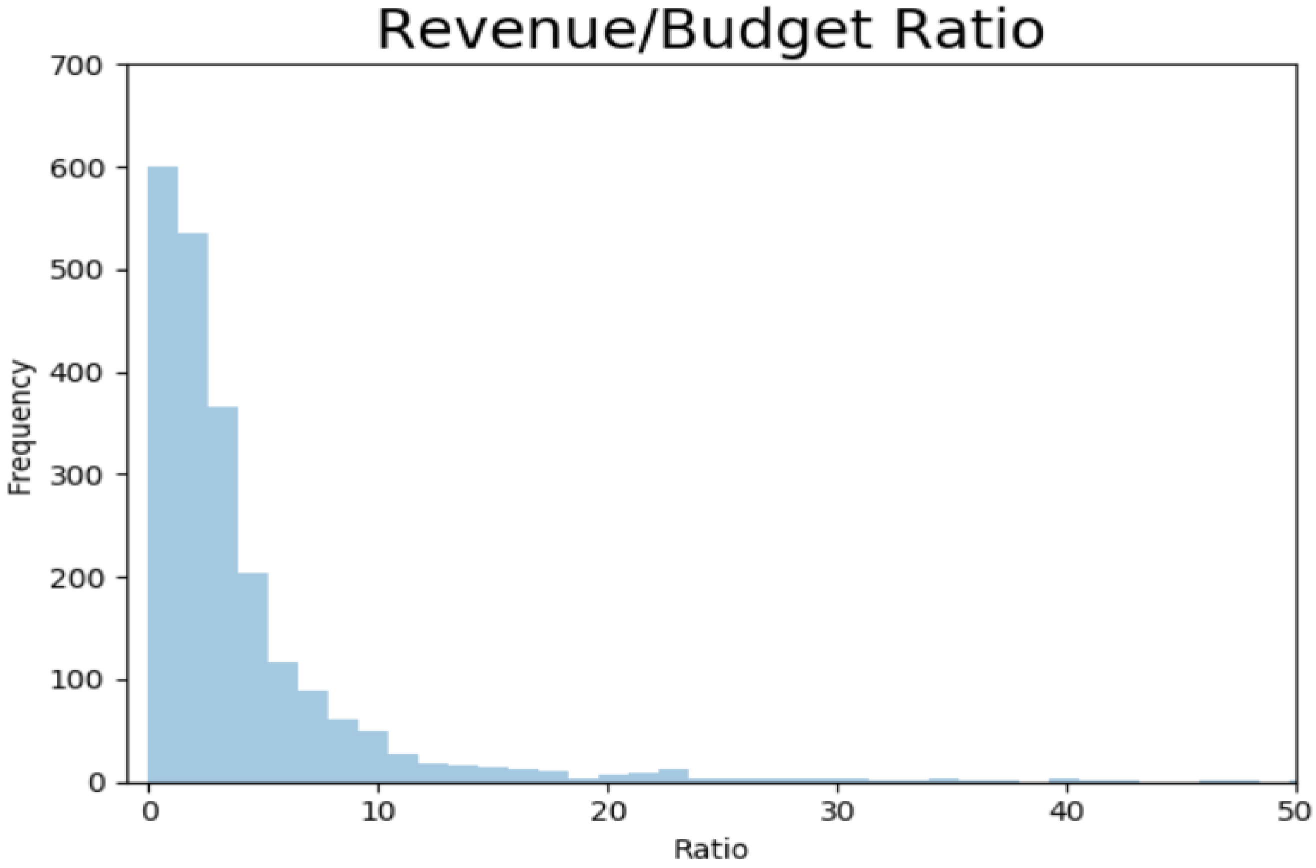

Movie Count by Genre
Frequency
500
400
300
200
100
0
Animation
Adventure
Action
Western
Drama
Crime
Thriller
Horror
Science Fiction
Comedy
Romance
Mystery
Family
Fantasy
History
War
Music
Documentary
Genre

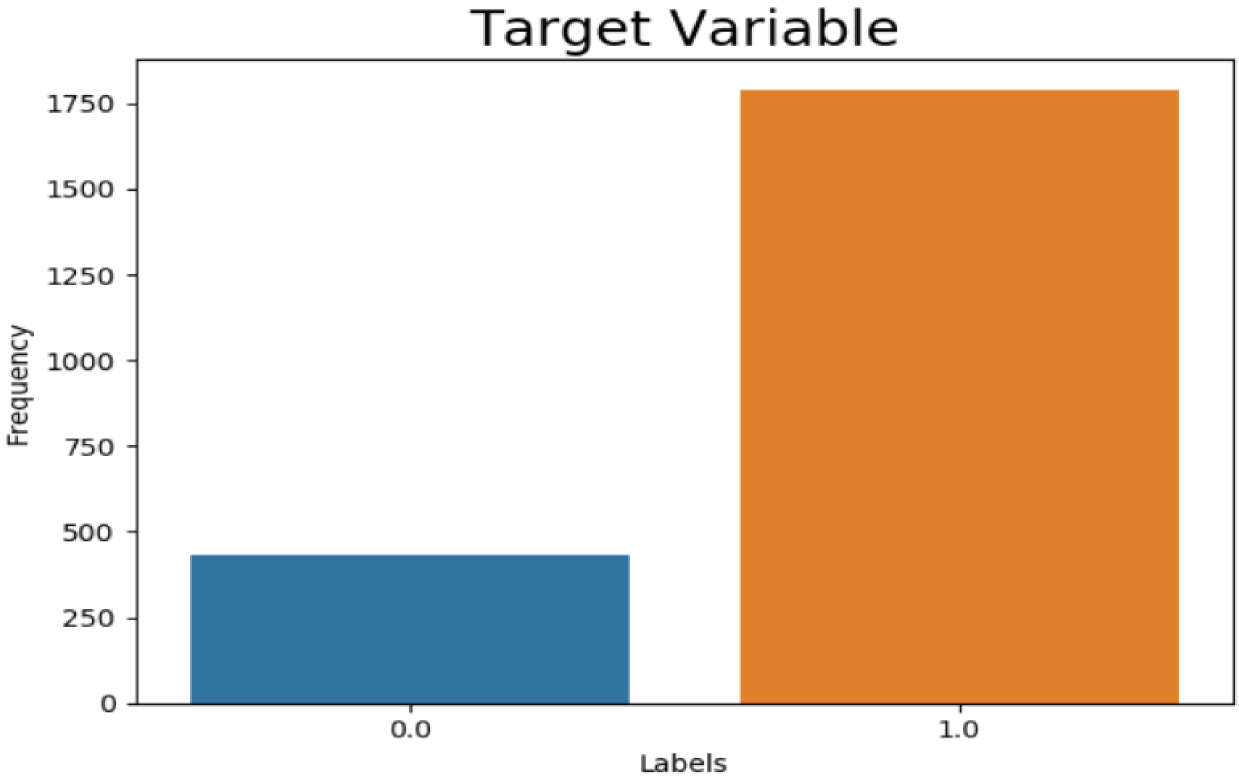

Movies by Release month
count
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
release_month


Top 10 Production companies
count
200
150
100
50
0
Paramount Pictures
Universal Pictures
Twentieth Century Fox Film Corporation
Columbia Pictures
New Line Cinema
United Artists
Miramax Films
Columbia Pictures Corporation
Warner Bros.
Walt Disney Pictures
Production_Company

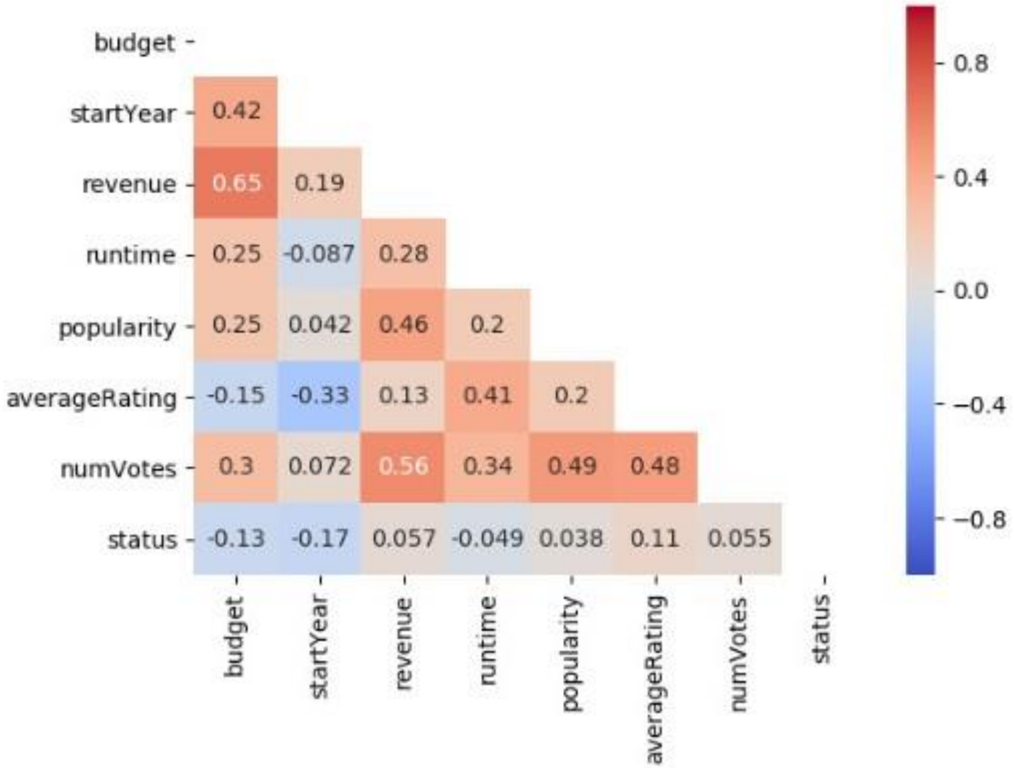

Movie Runtime
Frequency
0
50
100
150
200
75
100
125
150
175
200
225
250
Duration in Minutes


Movie Rating
Frequency
0
25
50
75
100
125
150
175
2
3
4
5
6
7
8
9
Movie Rating


Revenue/Budget Ratio
Frequency
700
600
500
400
300
200
100
0
0
10
20
30
40
50
Ratio


Target Variable
Frequency
1750
1500
1250
1000
750
500
250
0
0.0
1.0
Labels

Correlation Heatmap



	budget	startYear	revenue	runtime	popularity	averageRating	numVotes	status
budget								
startYear	0.42							
revenue	0.65	0.19						
runtime	0.25	-0.087	0.28					
popularity	0.25	0.042	0.46	0.2				
averageRating	-0.15	-0.33	0.13	0.41	0.2			
numVotes	0.3	0.072	0.56	0.34	0.49	0.48		
status	-0.13	-0.17	0.057	-0.049	0.038	0.11	0.055	

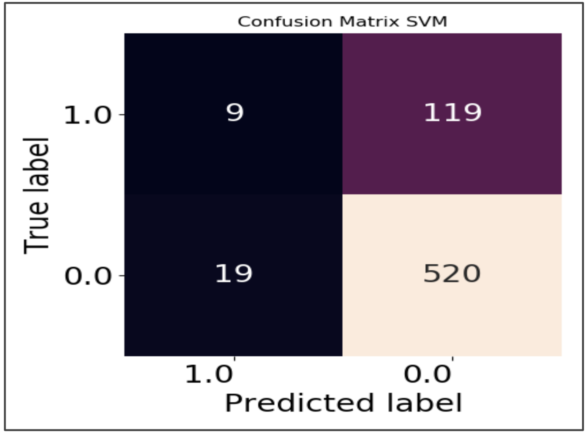
Modeling

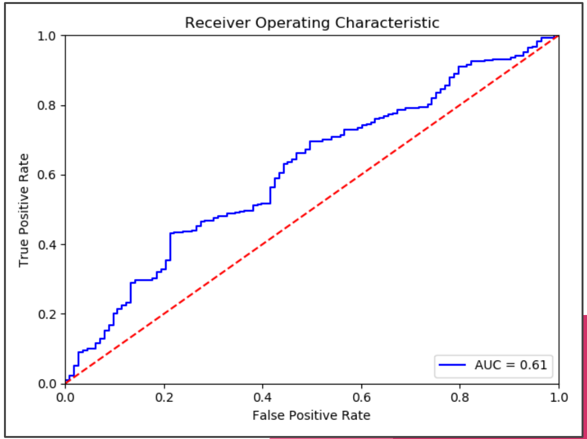
1. Data Splitting
2. Stratified Sampling
3. Oversampling
4. Label Encoding
5. Scaling
6. Feature selection

Continued...

Results with four features:

Model	Accuracy
Accuracy DT Entropy	69.26%
Accuracy SVM	81.40%
Accuracy RF	80.35%
Accuracy KNN	71.81%
Accuracy NB	22.93%

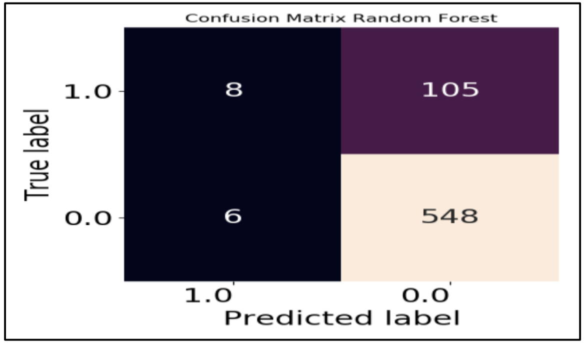


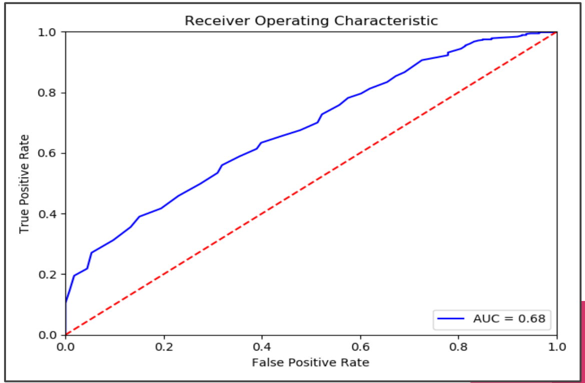


Area Under the Curve = 61%

Continued...

Model	Accuracy
Accuracy DT Gini	73.91%
Accuracy SVM	81.41%
Accuracy RF	83.35%

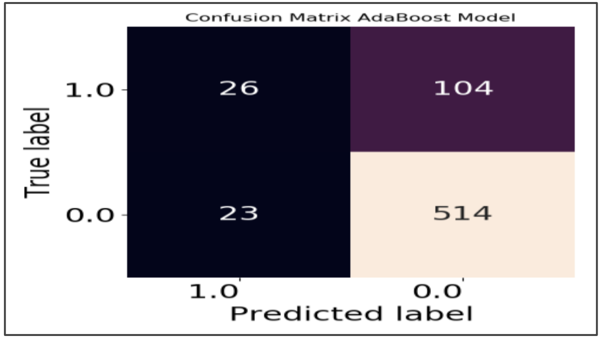


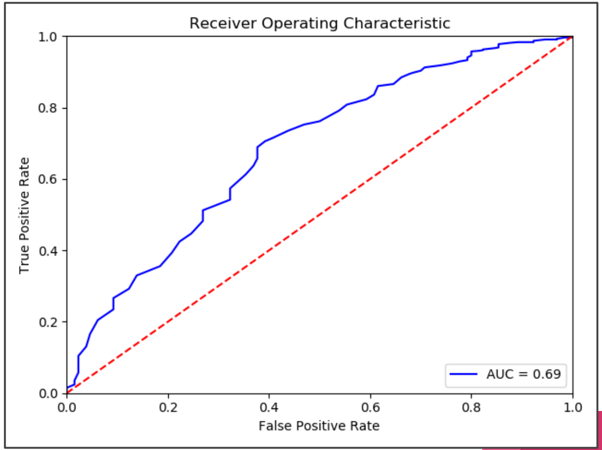


Area Under the Curve = 68%

Continued...

Model	Accuracy
Decision Tree (Entropy)	72.56%
Support Vector Machine	69.12%
Random Forest	79.91%
Bagging(Mode)	78.41%
Adaptive Bootstrap	80.96%





Area Under the Curve = 69%
K = 20.57%

Limitation

- Biased Label
- Missing Values
- Invalid Data
- Additional Features

References

https://www.the-numbers.com/market/

http://www.diva-portal.org/smash/get/diva2:1106715/FULLTEXT01.pdf

https://io9.gizmodo.com/how-much-money-does-a-movie-need-to-make-to-be-profitab-5747305

Maklin, C. (2019). *AdaBoost Classifier Example In Python*. [online] Medium. Available at: https://towardsdatascience.com/machine-learning-part-17-boosting-algorithms-adaboost-in-python-d00faac6c464