Movie Success Prediction

Team 1:
Abhishek Gupta | Ajith Hegde | Gabriela Caballero | Palak Jain | Russ Kaehler

NYU TANDON SCHOOL OF ENGINEERING

Project Overview

1. Problem Analysis

2. EDA

3. Data Enhancement & Normalization

4. Feature Engineering

5. Building Models and predictions

6. Recommendations

Problem Statement

$35BN

Movie Industry profit in 2019 in the US.

30%

of it comes from movie theatres and rest from different sources.

Increasing trend of streaming services that creates the need of more information and advice for making the right financial choices.

Data Sources

THE MOVIE DB

87k Records
From 176 Countries

2017 - 2021
Years

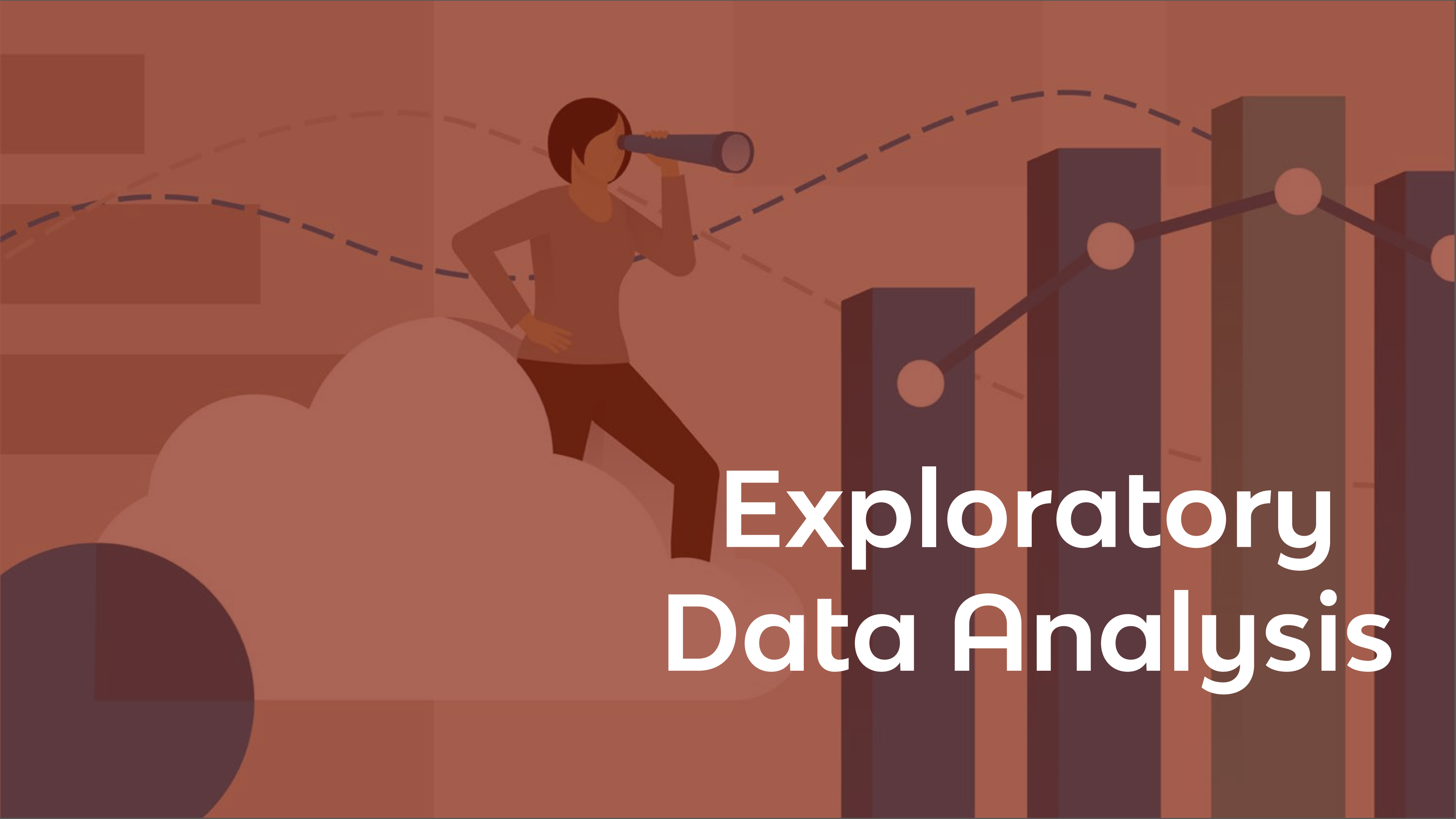
36 Columns
Numerical , Binomial, Text data types

IMDb

6.4k Records
Based on Final Cleaned Data of TMDB records

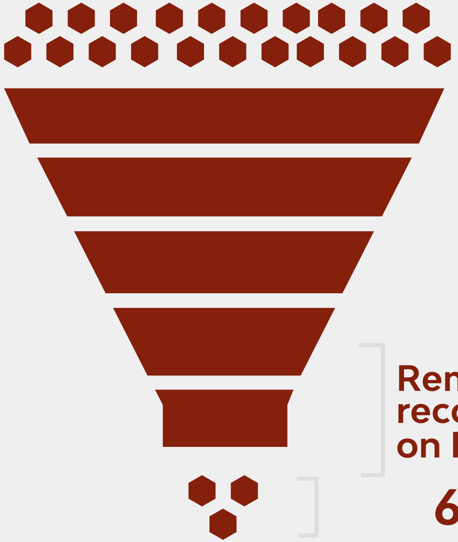
2017 - 2020
Years

3 Columns
Ratings and Vote counts


Exploratory
Data Analysis

Data Cleaning

Imputation for Numerical Features & Dummy Variable Creation for Categorical Features



87k Records

Minus 75 countries with no data on Revenue & Budget

34k Records

Remove all remaining records with no data on Revenue & Budget

6.4k records

6.4k Movies from 2017-20

19
Genres

$40B
Budget

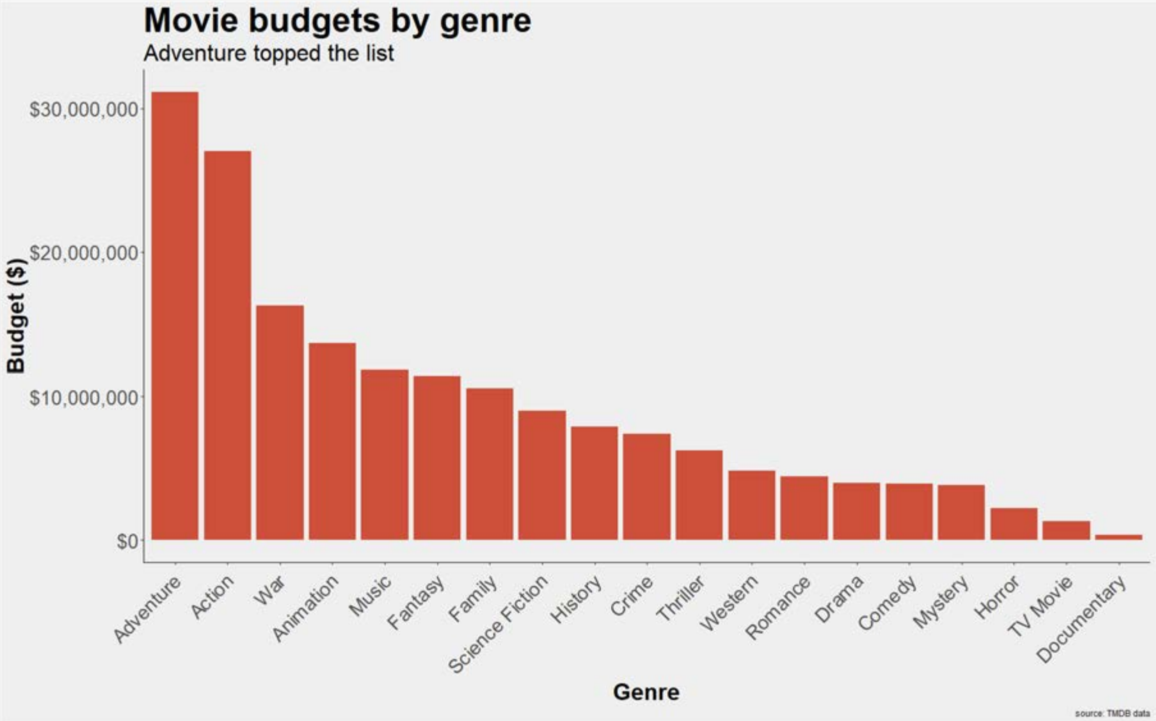
$337B
Revenue

6.4
Avg. rating

81 min
Avg. runtime

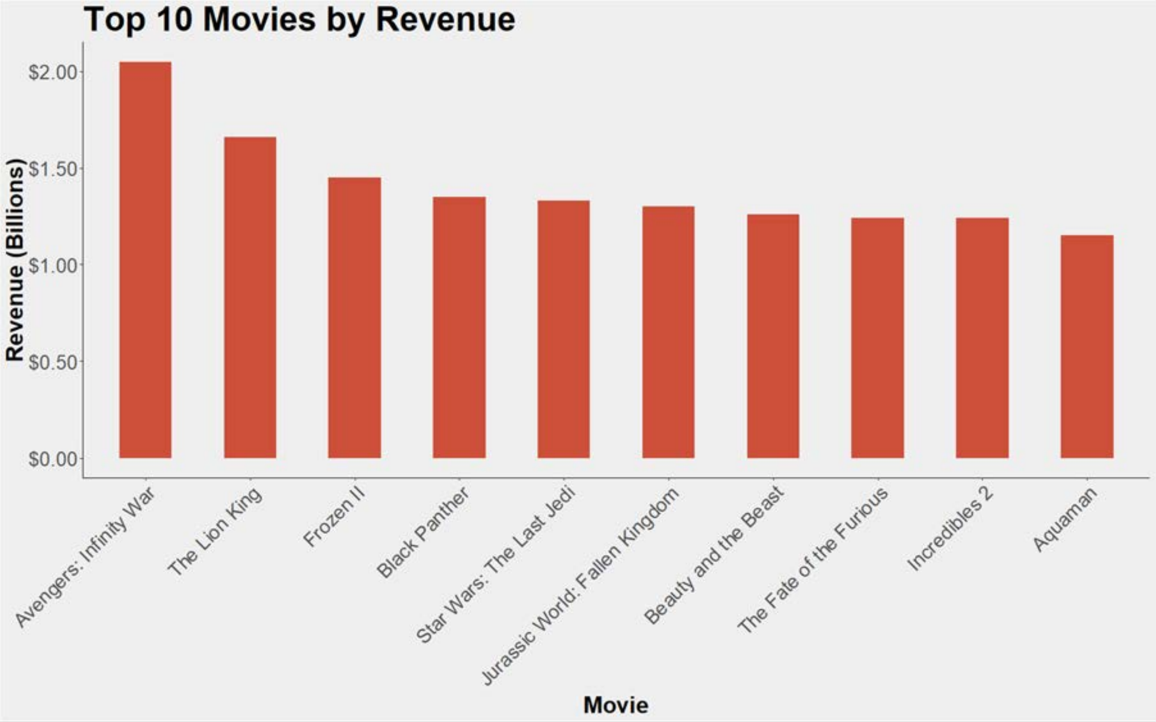
111
Countries

EDA



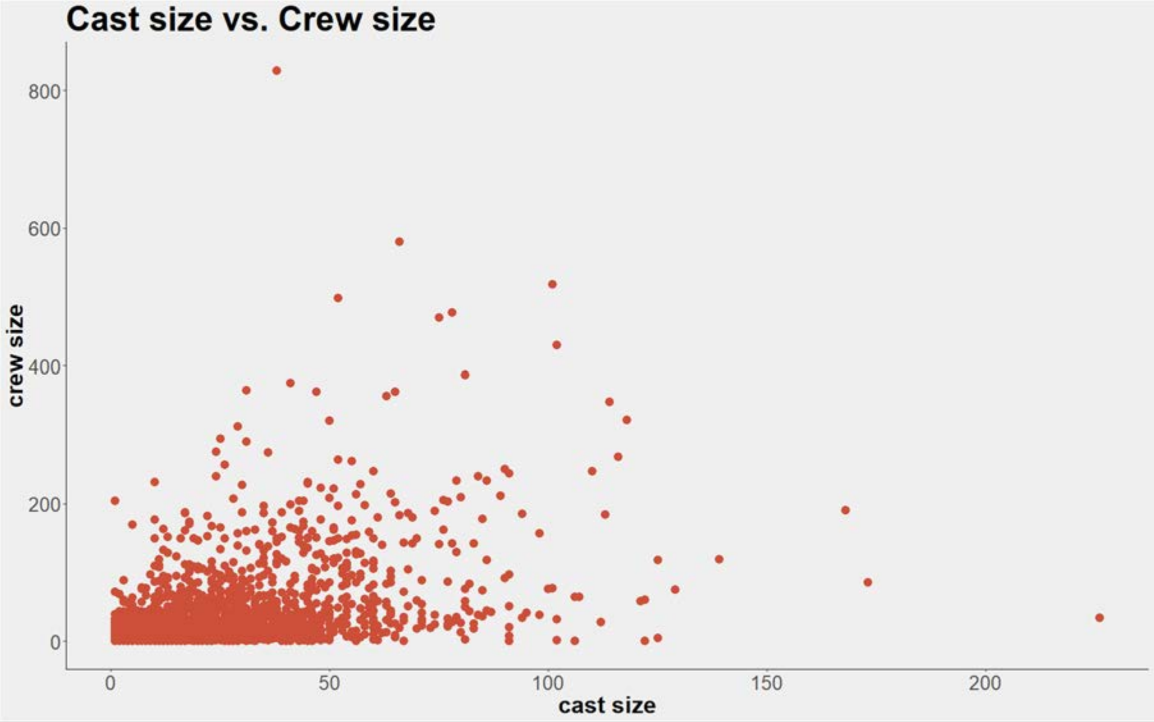
Movies in ‘adventure’ genre have the highest budget of more than $30M, followed by action and war.

EDA



Avengers and Lion king were the mega hits in theatres worldwide with gross revenue of over $1.5B

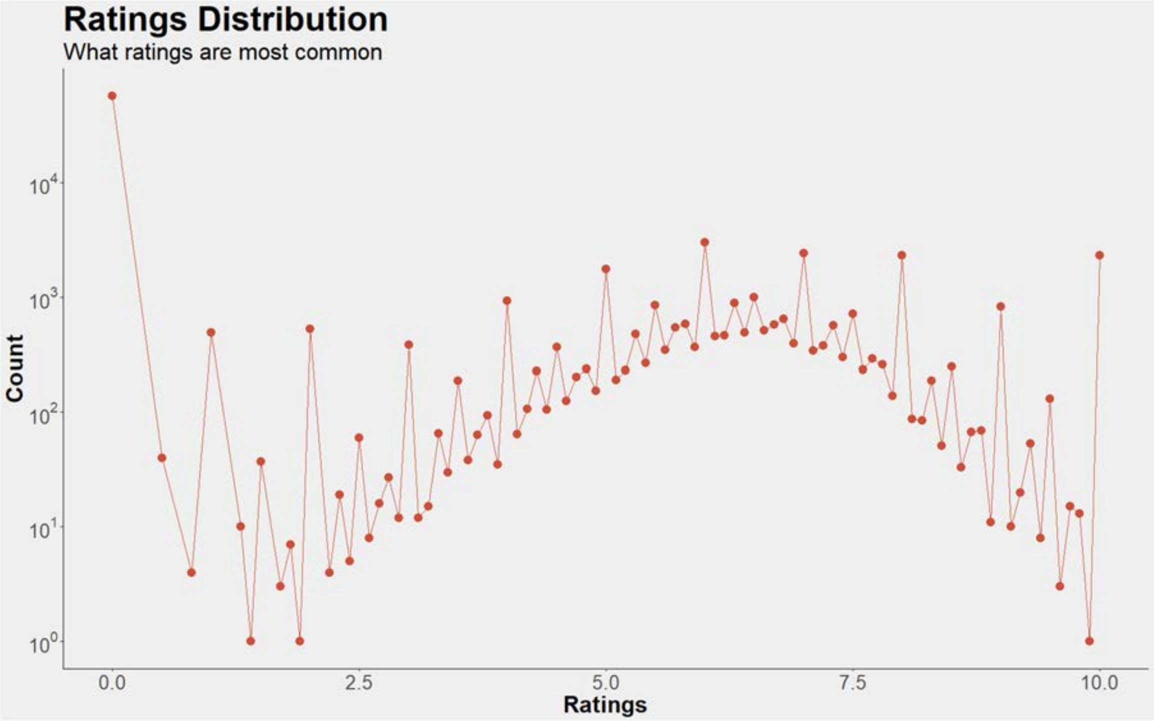
EDA



Crew size is generally bigger than the Cast size.

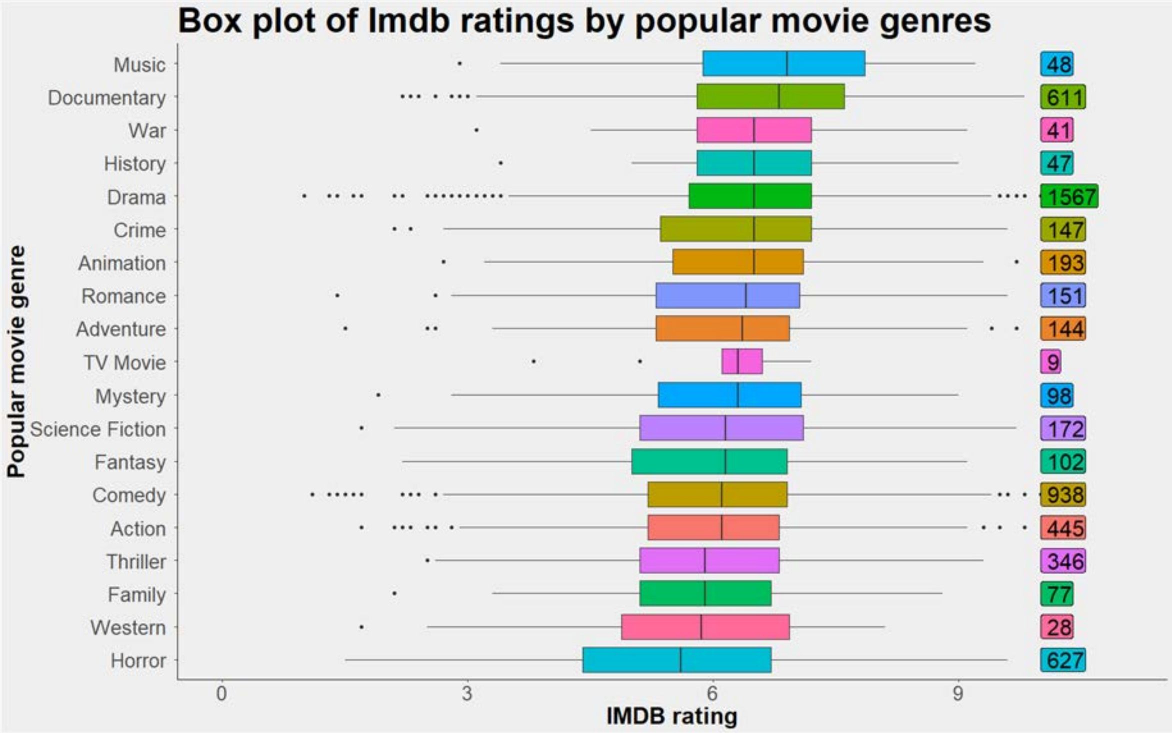
Cast size upto 50 and Crew size upto 200 covers most movies.

EDA



Users ratings from 5 to 7 are the most common for majority of the movies.

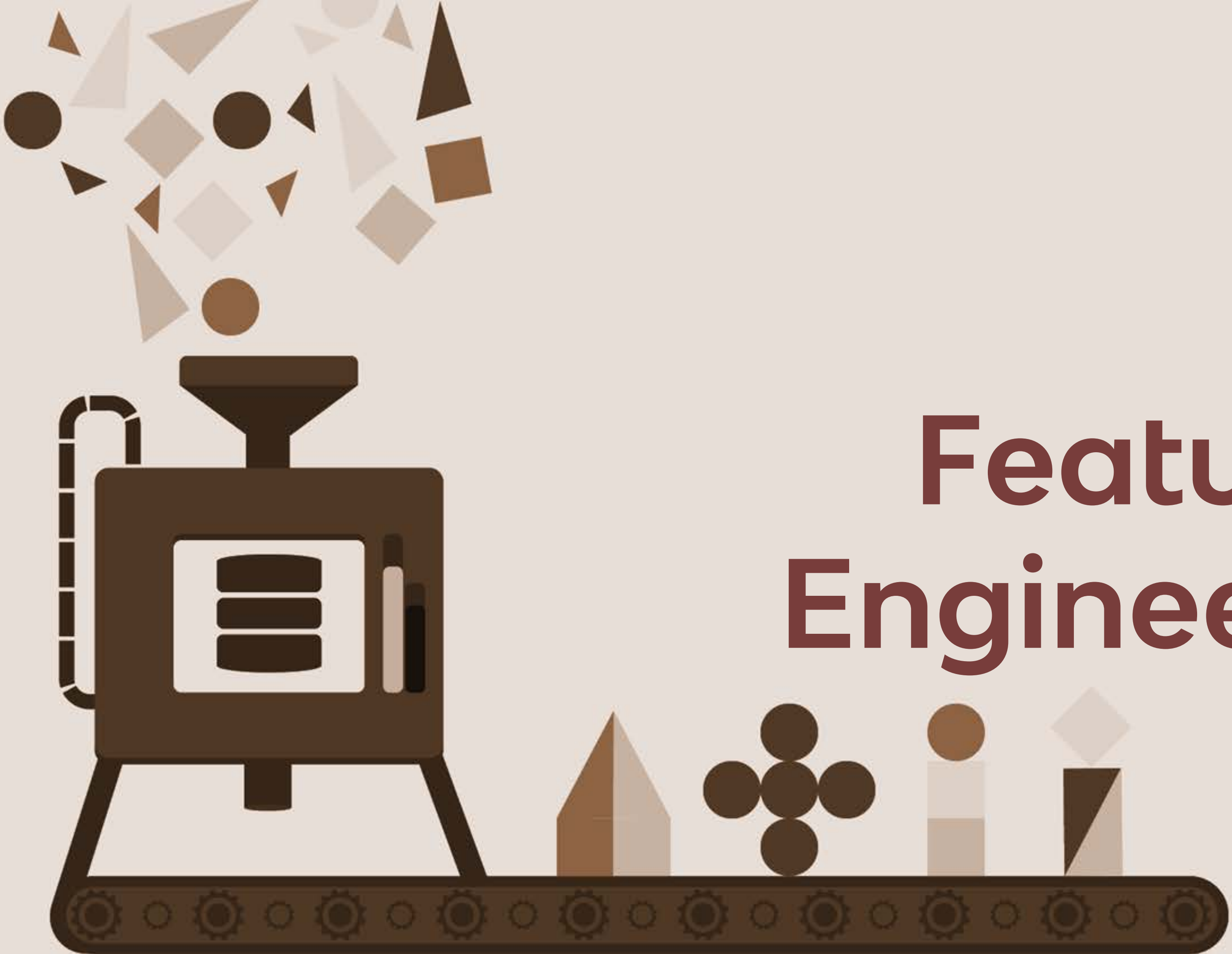
EDA



Action, Adventure, and War movies that are highly budgeted gets slightly lower Imdb ratings than Documentaries and Movies.

Music has relatively higher ratings, having a median score of approx 7 but only have 48 movies.

Horror movies receives lower ratings with median of < 6.



Feature Engineering

Inner Join to add IMDB Ratings and Vote Counts

Converted 11 Text Features like genres, keywords, etc. into 13 numerical and binomial features

Normalized Profitability and based classification on it

*Profitability= ((Revenue - Budget)/ Budget) * 100*

11 Features created

- Movie directed by a top 10 director
- Movie produced by a profitable production company
- Avg. revenue per movie in a collection
- Competition during release (# of movies released in the same week)
- Cost per Capita (Budget /(Cast size + Crew Size)
- Within 1-σ Combined Rating

Modeling

Modeling

25+ New Features

Numerical and Binomial

50 features

Off 513 dummified columns, 50 Top Important Features were selected

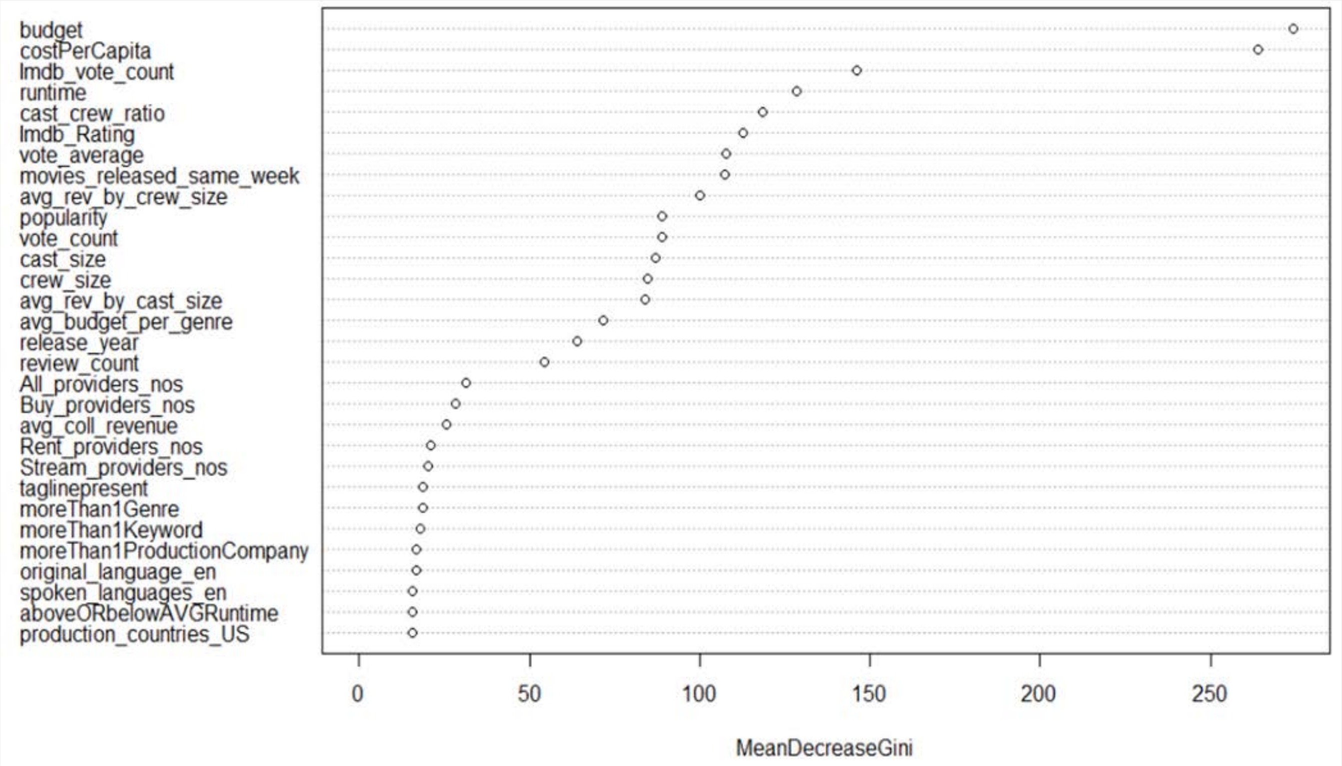
ROSE Balancing

Both Under & over Sampling is used

4 Analytical Models

2 Glm and 2 Random Forest

Feature Selection



Random Forest is used for Feature Selection

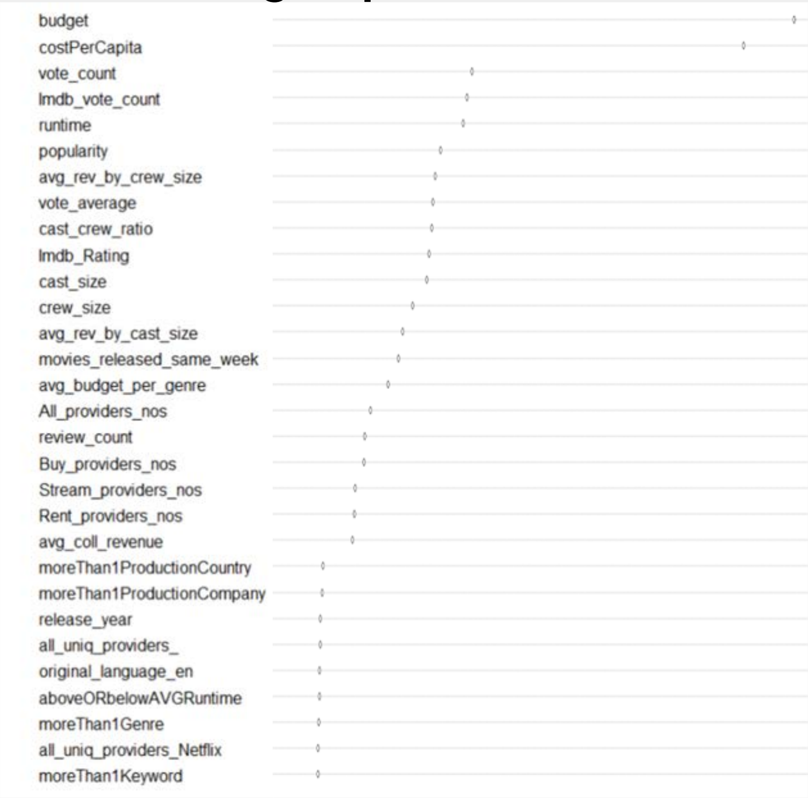
Top 50 features are taken from 513 features for further Model creation

Comparing Models

	Data Split	Accuracy	Precision	Recall	F-Score	Sensitivity	Specificity
GLM 1	Random 80:20	65%	0.68	0.77	0.72	0.78	0.46
RF 1	Random 80:20	90%	0.90	0.92	0.91	0.92	0.86
GLM 2	Train: 2017-18 Test: 2019	65%	0.67	0.82	0.74	0.82	0.38
RF 2	Train: 2017-18 Test: 2019	91%	0.90	0.95	0.92	0.95	0.83

RF2

Features by Importance

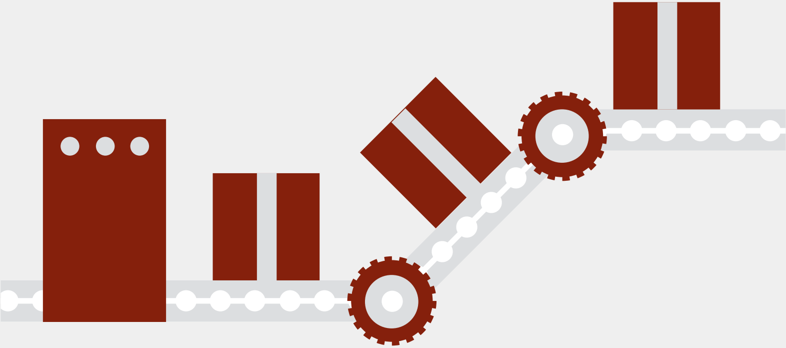


Confusion Matrix & Statistics

```
          Reference
Prediction   0   1
         0 721  83
         1  41 417

               Accuracy : 0.9017
                 95% CI : (0.884, 0.9176)
    No Information Rate : 0.6038
    P-Value [Acc > NIR] : < 0.0000000000000022

                  Kappa : 0.7916

 Mcnemar's Test P-Value : 0.0002315

            Sensitivity : 0.9462
            Specificity : 0.8340
         Pos Pred Value : 0.8968
         Neg Pred Value : 0.9105
             Prevalence : 0.6038
         Detection Rate : 0.5713
   Detection Prevalence : 0.6371
      Balanced Accuracy : 0.8901

       'Positive' Class : 0

Area under the curve (AUC): 0.890
```

Text Analysis

Text Processing



1. Stopwords filtered
2. Stemming
3. Words in context analysis
4. Sentiment analysis

Successful movies

Overview

Keywords

Successful Movies overview and keywords in Context

Mysterious films

Drama movies involving deaths and CIA

Comedies around family and friends

Sentiment Analysis dictionaries used from the tidytext package

1. Bing Dictionary from Bing Liu and collaborators

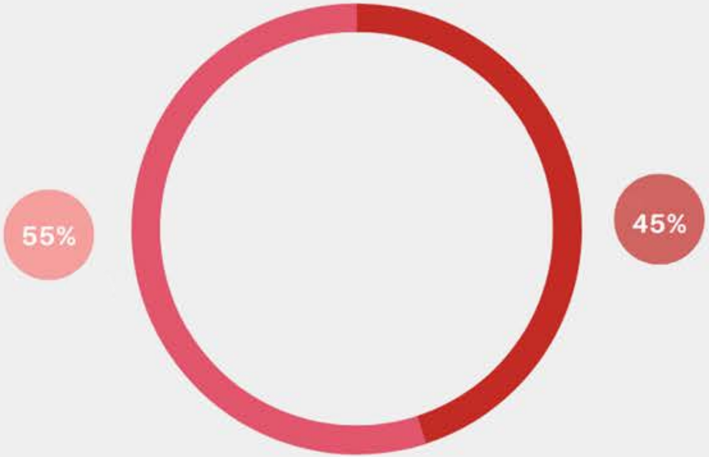
2. AFINN from Finn Årup Nielsen

3. NRC from Saif Mohammad and Peter Turney

Successful movies

Sentiment Analysis on Reviews

Positive Reviews



Negative Reviews

Unsuccessful movies

Overview

Keywords

Unsuccessful Movies overview and keywords in Context

Book based and biography films

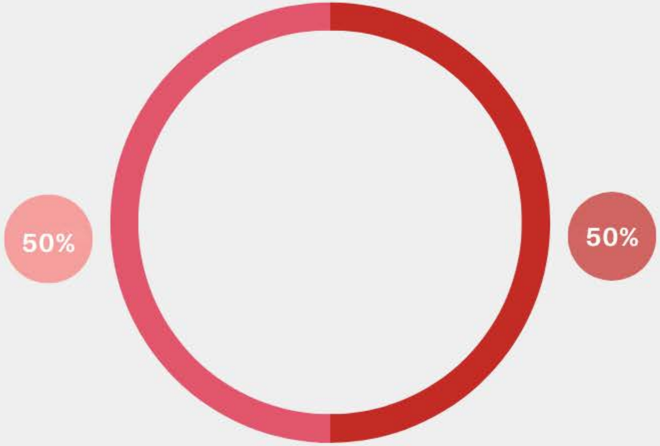
War and violence

Based on minorities (black, lgtb, female)

Unsuccessful movies

Sentiment Analysis on Reviews

Positive Reviews



Negative Reviews

RECOMMENDATIONS

Budget & Cost Per Capita

Budget **lower** than **Avg. Budget per Genre** and with **smaller crew+Cast**
High Cost/Capita keeps everyone motivated

Online Availability - Netflix

Accessible through more providers in rent, **buy & stream** options | **Avg.: 4**
Movies on Netflix perform better in terms of profit

Ideal Runtime for Feature films

Keep Runtime between **47-128** minutes
Within 1σ of Median Runtime

Production Company

Have **more than 1** Production Company
Better Network, Budget & skill

Rating & Popularity

Vote counts matter more | Popularity
Both IMDB & TMDB **ratings**: between **7-9**

Collection & Tagline

Movies part of a collection **perform better. Taglines** can also help increase **profitability**

Competition

Lower competition is better
Choose a week with less than **18** releases globally.

Production Country & Language

With **English** as original language, aim to **produce** in **more than 1 country** for better results

The End!

or the beginning 🙂