



Streaming Platform Analysis

Group 6

CHEN, Taoyue 1155141543
LAU, Ho Cheung 1155157540
LI, Yuji 1155157174

POON, Yin Ki
ZHAO, Shengchun 1155157253



Problem Statement

- ★ How to choose a streaming platform?
- ★ Should audiences pay for newly released movie ?

Introduction of the Dataset



Two sets of dataset



Part 1 & 3

16 thousand movie records and
their features



Part 2

movie records and their release
date on the corresponding
platform

Approach to the Problem

Part 1

Expository data analysis of our movie data to navigate platform-based characteristics of movies

Part 2

Forecasting the number of new movies available on each platform(Hulu, Disney, Netflix)

Part 3

Modelling the relationship between features of a movie and its IMDb rating

The background is a stylized stage scene. It features a dark blue backdrop with a spotlight beam shining from the bottom left onto a red carpeted area. On the right side of the stage is a red podium with two microphones. In the foreground, there are yellow stanchions connected by red ropes. The entire scene is framed by red curtains on the left and right sides.

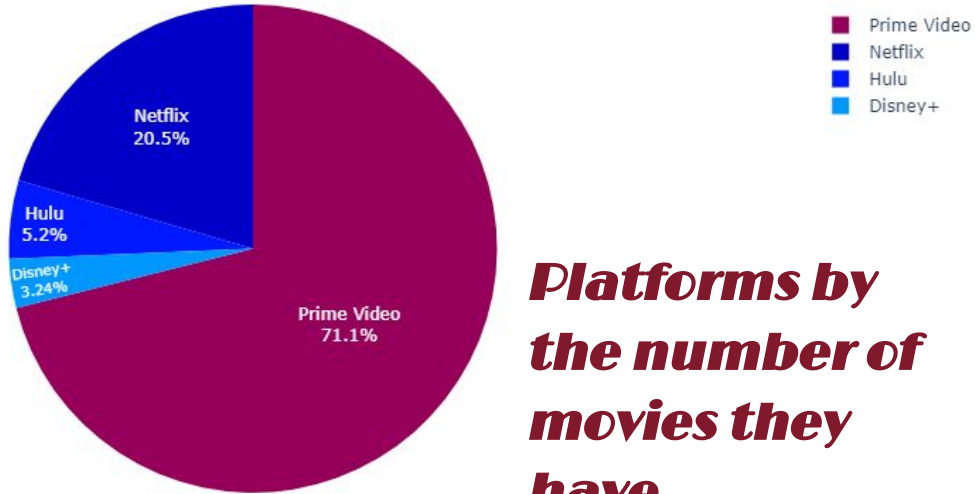
01

Expository Data Analysis result

Analysis result in case 25

Movie Count Of Different Platforms

01



***Platforms by
the number of
movies they
have***

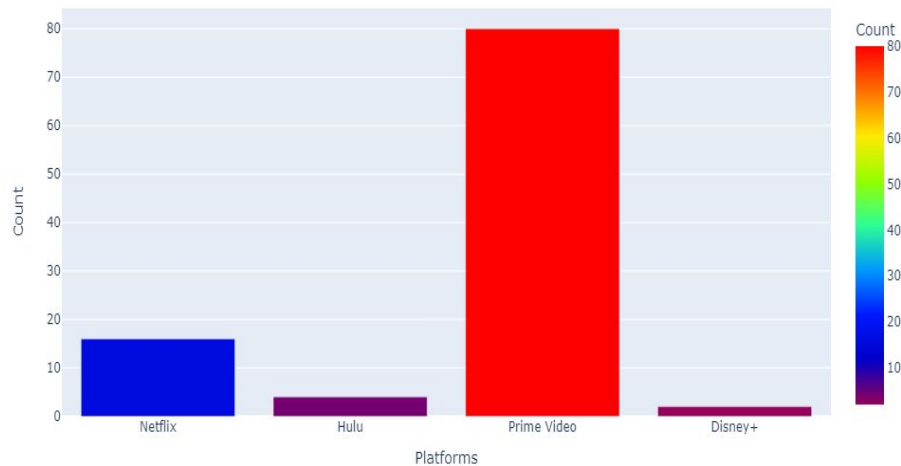
Analysis result in case 25

02

Platforms by the number and percentage of high-quality(IMDb 8.5+) movie

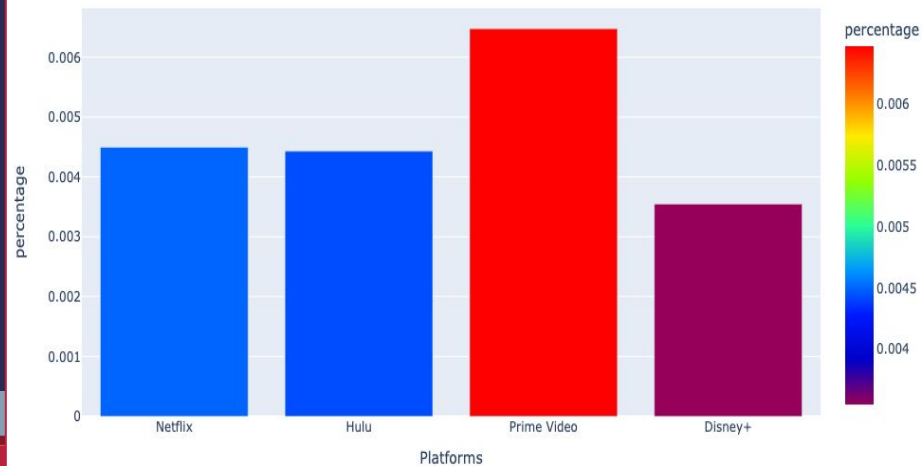
By quantity:

IMDB 8.5+ Movies on different Platforms



By percentage:

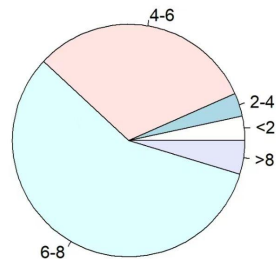
percentage of IMDB 8.5+ Movies on different Platforms



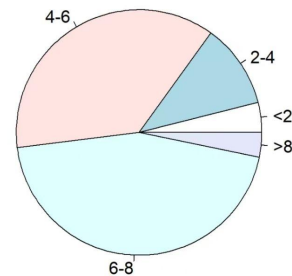
02

*Platforms by the
number and
percentage of
high-quality(IMDb
8.5+) movie*

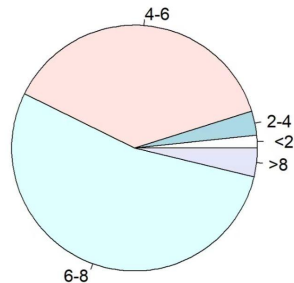
Rating Distribution of Movies in Netflix



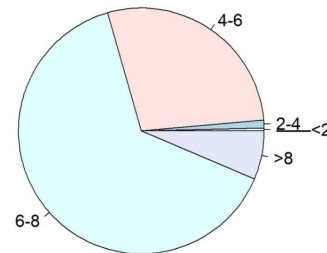
Rating Distribution of Movies in Prime Video



Rating Distribution of Movies in Hulu



Rating Distribution of Movies in Disney+



Extended Analysis of case 25

Movie Count By produced Year across platform

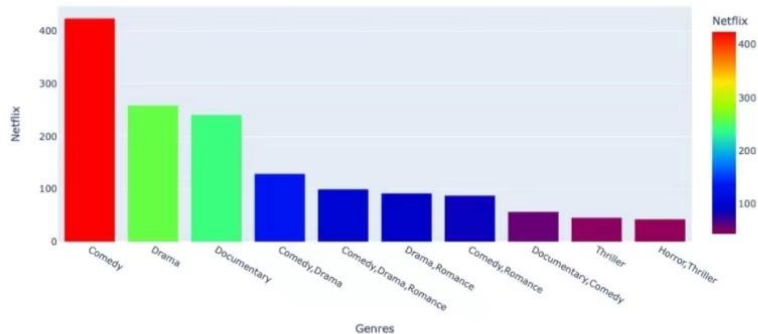


Extended Analysis of case 25

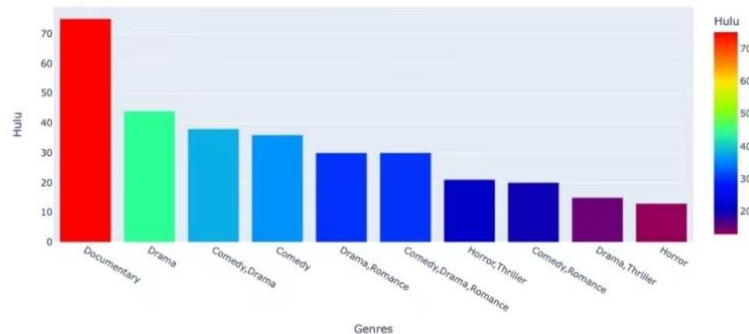
04

Platforms by their top ten genres of movies

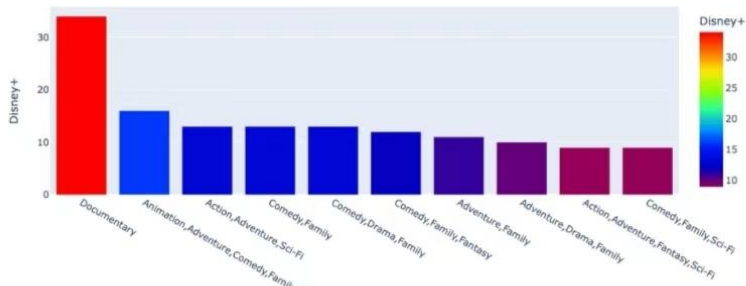
Top 10 Genres Movie Count on Netflix



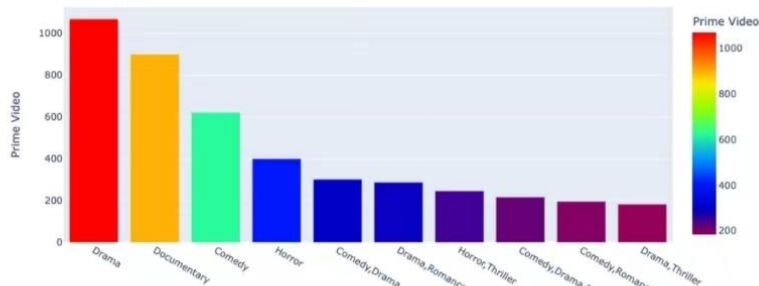
Top 10 Genres Movie Count on Hulu



Top 10 Genres Movie Count on Disney+



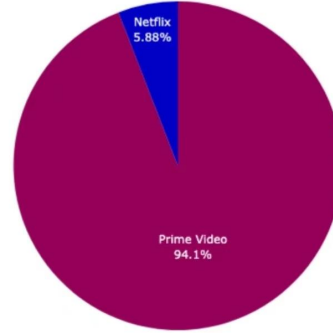
Top 10 Genres Movie Count on Prime Video



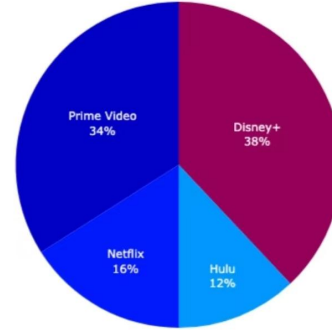
Extended Analysis of case 25

05

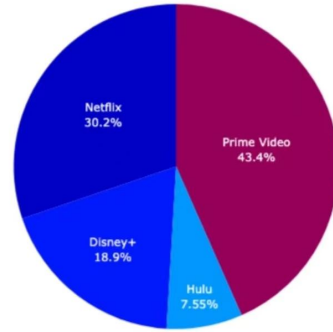
*Platforms shares of top 50
movies in **Genre**
Documentaries, Action,
Animation, and
Adventure*



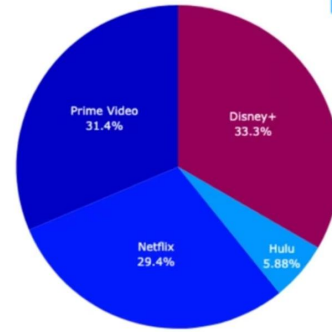
Documentaries



Animation



Action



Adventure



Extended Analysis of case 25

90%

Netflix

95%

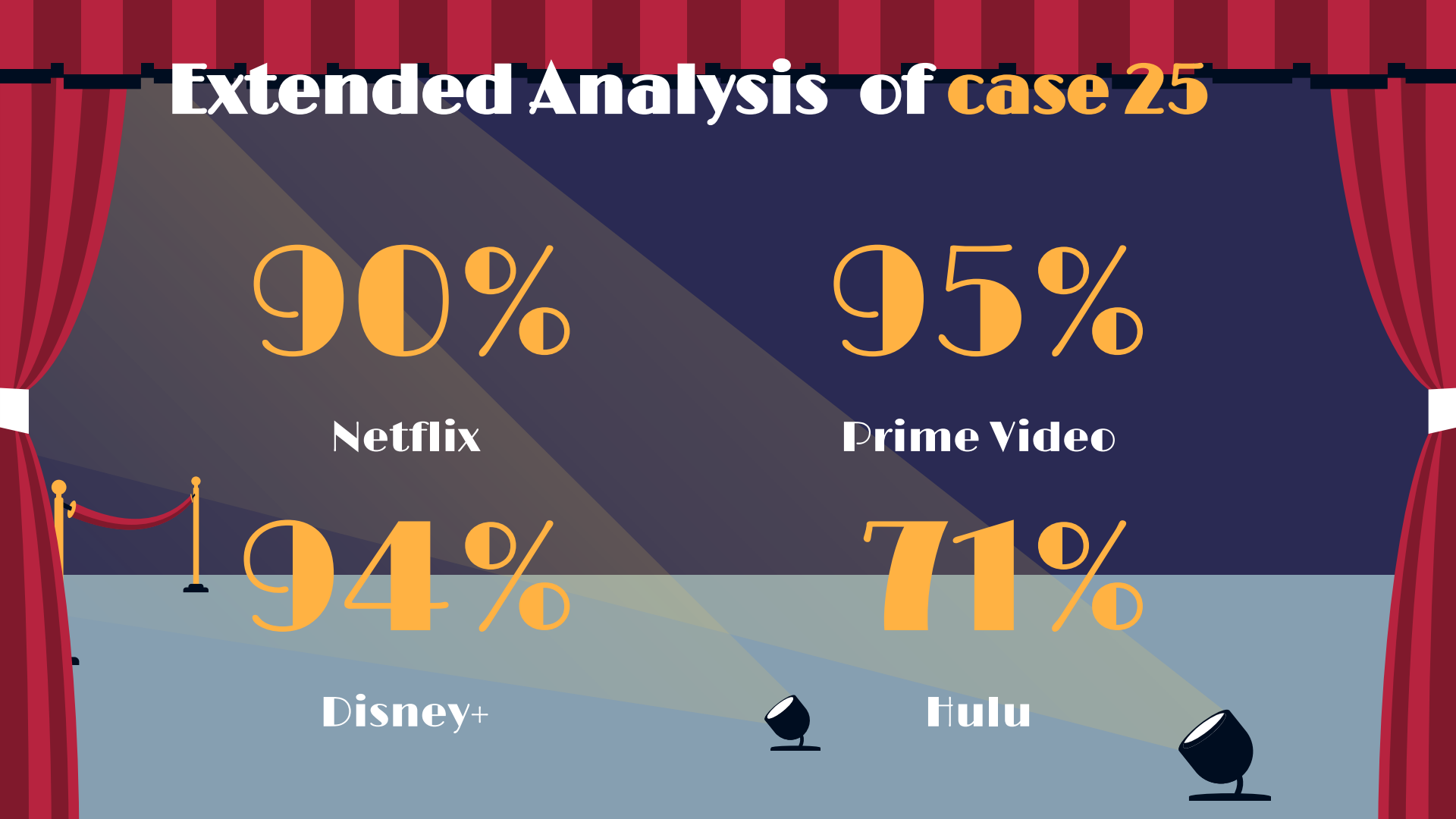
Prime Video

94%

Disney+

71%

Hulu



Limitation ?

Completeness

or

representability
of our dataset



An illustration of a hand holding a clapperboard, with a spotlight beam shining on the text area. The background is dark blue with red curtains on the sides.

02

Forecasting the number of new movies in 2023

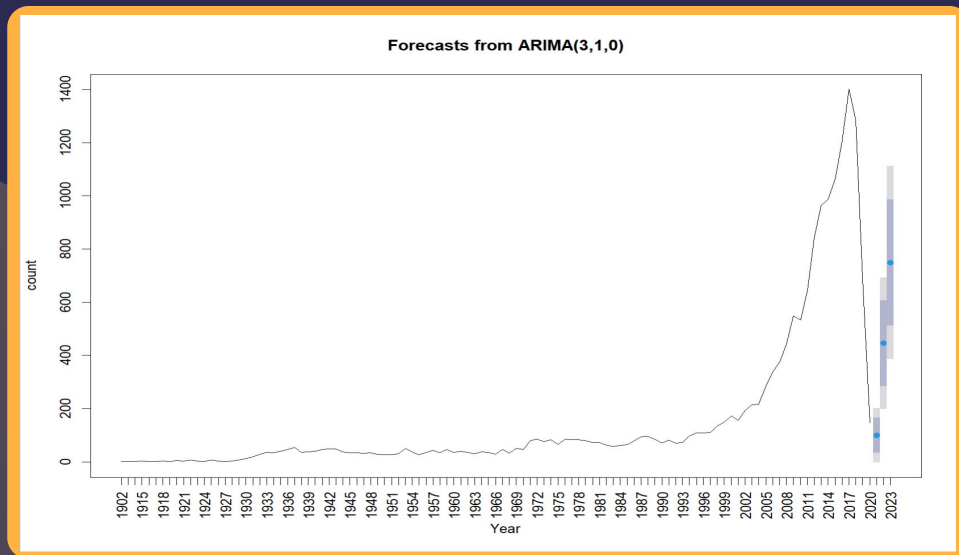
Which platforms will have
more new contents?



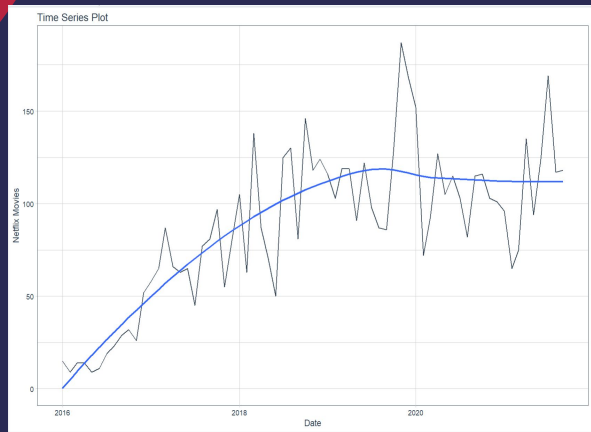
Overall Movie Market Forecasting

Rapidly developed around 2002

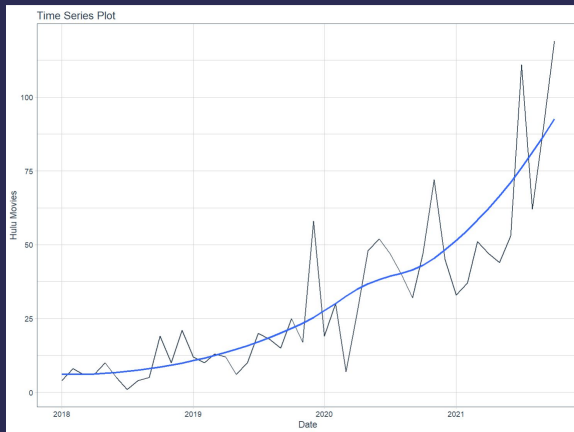
Reached the peak around 2017



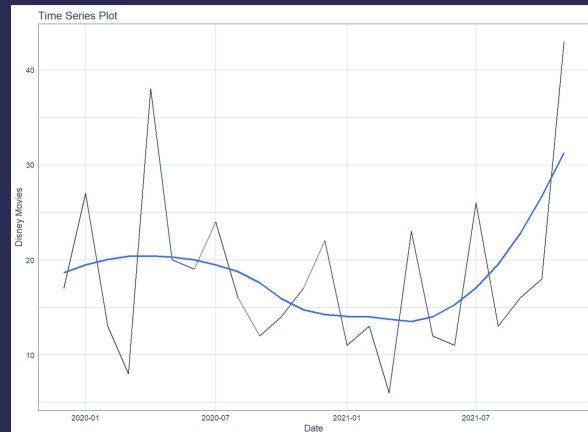
Data Visualization



Netflix



Hulu



Disney

Seasonality Testing



Netflix



Hulu



Disney

Forecasting The Number Of New Movies

Will Be Released On Different Streaming Platforms



Netflix



Disney



Hulu



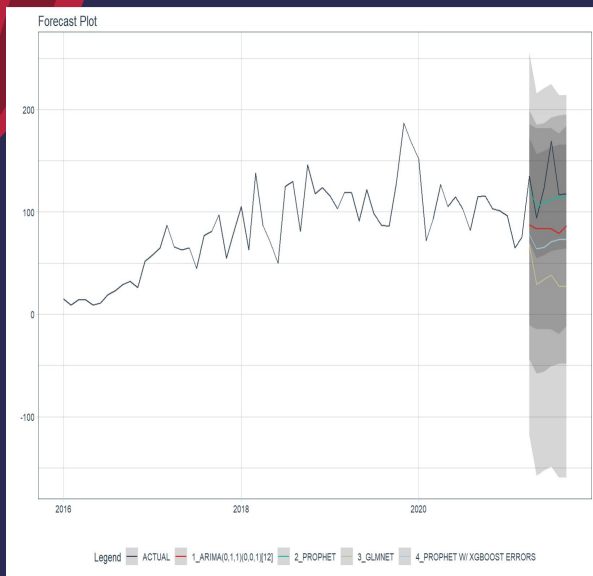
Model We Used

- 1, Auto-ARIMA Model
- 2, Prophet Model
- 3, Elastic Net (GLMNET)
- 4, Hybrid ML Model
(combine Prophet and XGBoost)

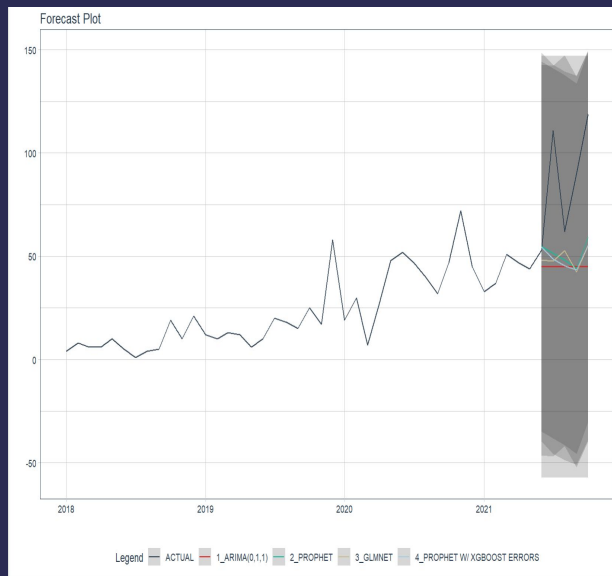




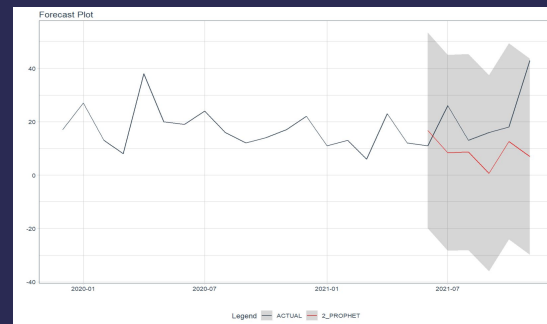
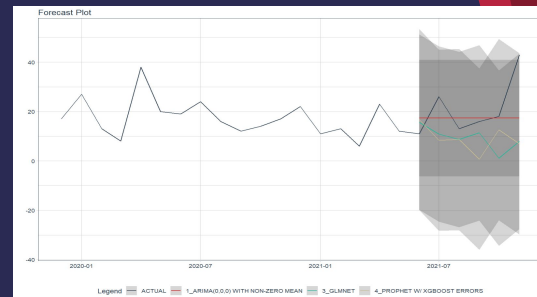
Validation



Netflix

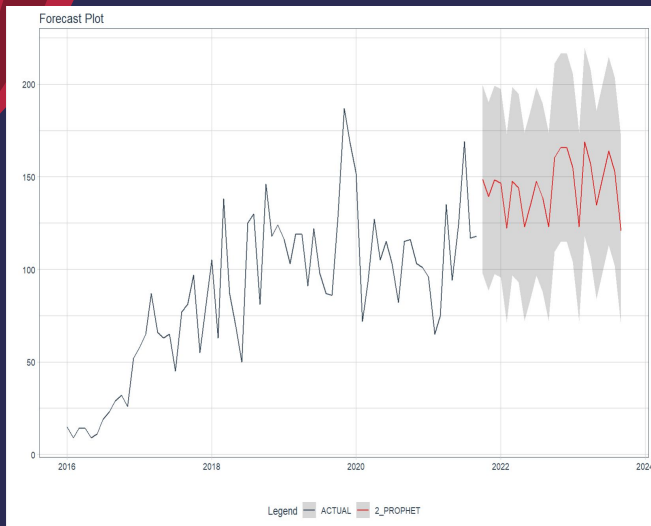


Hulu

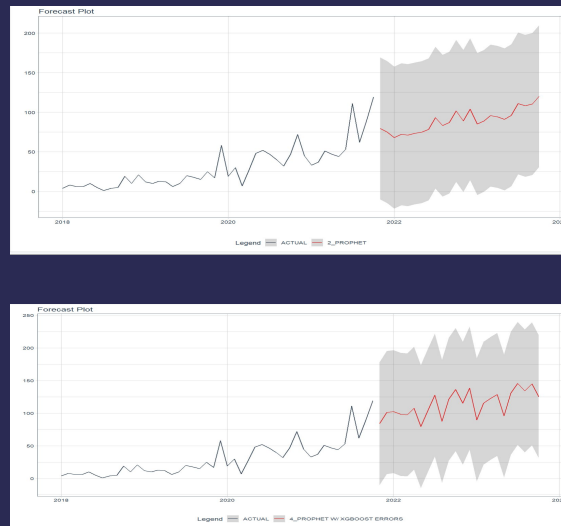


Disney

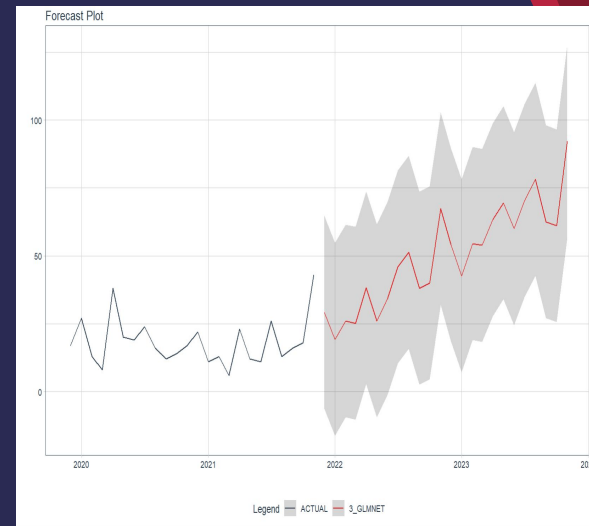
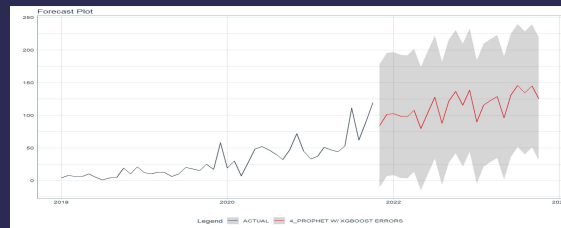
Forecasting For Each Platforms



Netflix



Hulu



Disney

Platform recommendation



★ Netflix



Limitations

- 1, Not add other variables to the model**
- 2, Some uncontrollable societal factors may have an influence to the forecasting result**
- 3, Hard to verify the completeness of the dataset**

An illustration of a hand holding a clapperboard, with a spotlight beam shining on it from the left. The clapperboard has a black top bar with white diagonal stripes, a black middle bar with white diagonal stripes, and a black bottom bar with white diagonal stripes. The hand is holding the clapperboard from the left side, with the thumb and index finger visible. The background is dark blue with red curtains on the left and right sides. A spotlight beam shines from the left, illuminating the clapperboard and the text on the right. A small black floor lamp is visible in the bottom right corner.

03

Model Development

How could we choose movies
on streaming platform ?

Which Movie is **worth watching** ?

- ★ The IMDb score can be used as a movie viewing guide.
- ★ But what about newly released movies that don't have rating?

We need to know what features of the movies affect the IMDb score.



Data Preprocessing

Standardization

Model requirements



Data cleaning

Remove & Fill the N/A



Dummy variables

For qualitative variables



The problems with OLS model

★ Over-fitting

train_R2	test_R2
0.360868406	-2.785E+13
train_RMSE	test_RMSE
1.0719	7133974.488
train_MAE	test_MAE
0.8314	184777.7621

Ridge regression model

Ridge

vs.

OLS

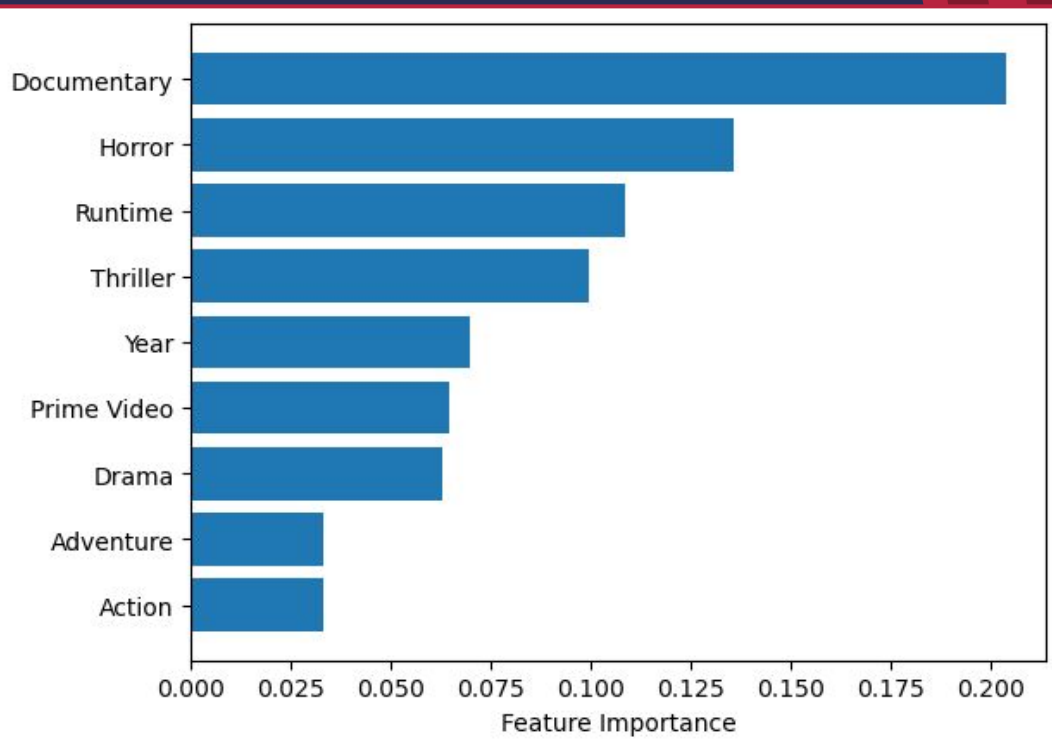


R2	train_R2	test_R2
Ridge	0.3528	0.3638
OLS	0.3609	-2.785E+13
RMSE	train_RMSE	test_RMSE
Ridge	1.0787	1.0782
OLS	1.0719	7133974.488
MAE	train_MAE	test_MAE
Ridge	0.8402	0.8374
OLS	0.8314	184777.7621

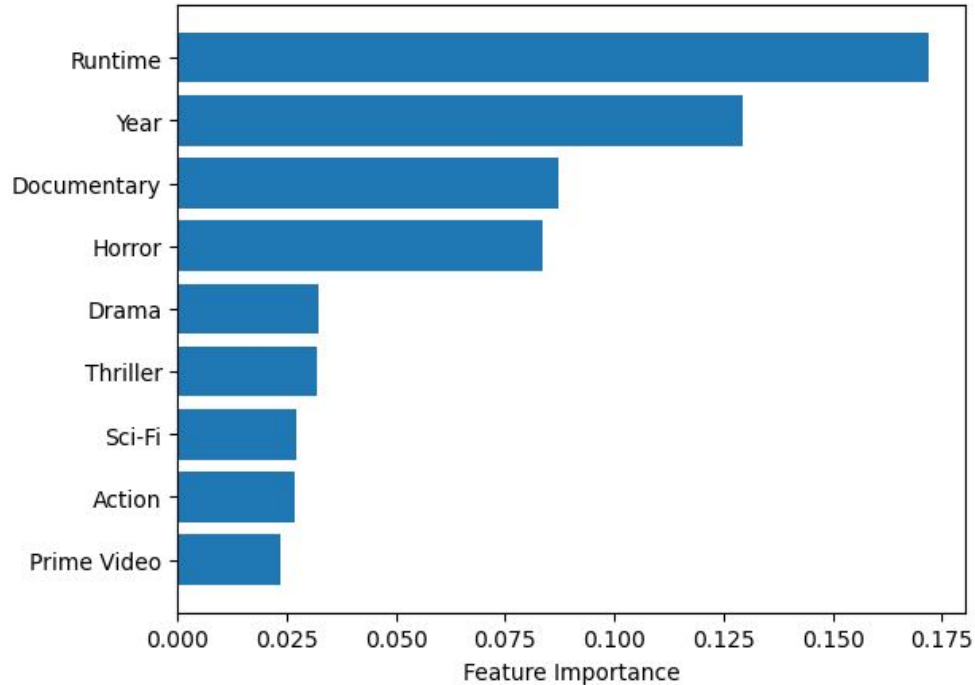
Decision tree

What features contribute the most to movie ratings?

R2	train_R2	test_R2
Decision Tree	0.4066	0.3358
RMSE	train_RMSE	test_RMSE
Decision Tree	1.0328	1.1017
MAE	train_MAE	test_MAE
Decision Tree	0.7983	0.851



Random forest



R2	train_R2	test_R2
Decision Tree	0.4066	0.3358
Random Forest	0.7731	0.437335321
RMSE	train_RMSE	test_RMSE
Decision Tree	1.0328	1.1017
Random Forest	0.6387	1.014
MAE	train_MAE	test_MAE
Decision Tree	0.7983	0.851
Random Forest	0.4739	0.7779

Model results

	train_RMSE	test_RMSE	train_MAE	test_MAE
OLS	1.0719	7133974.4877	0.8314	184777.7621
Lasso	1.0798	1.0790	0.8410	0.8382
Ridge	1.0787	1.0782	0.8402	0.8374
SVR	1.0808	1.0819	0.8270	0.8356
Decision Tree	1.0328	1.1017	0.7983	0.8510
Random Forest	0.6387	1.0140	0.4739	0.7779
Ada boost	0.4191	1.0557	0.2571	0.8023
Gradient boost	0.4295	1.2204	0.3311	0.9405

Limitations and application

- **Our weakness lies in need for other essential information about the movies.**
- **The model can be used as a helpful guide for audience to decide on whether to buy a movie online without previous reviews.**



Summary

- ★ **Exploratory data analysis for each platform**
- ★ **Time series and machine learning approaches to forecast the number of new movies in 2023**
- ★ **Model the relationship between movie features and IMDb ratings with regression and tree-based models**





Thank
you!