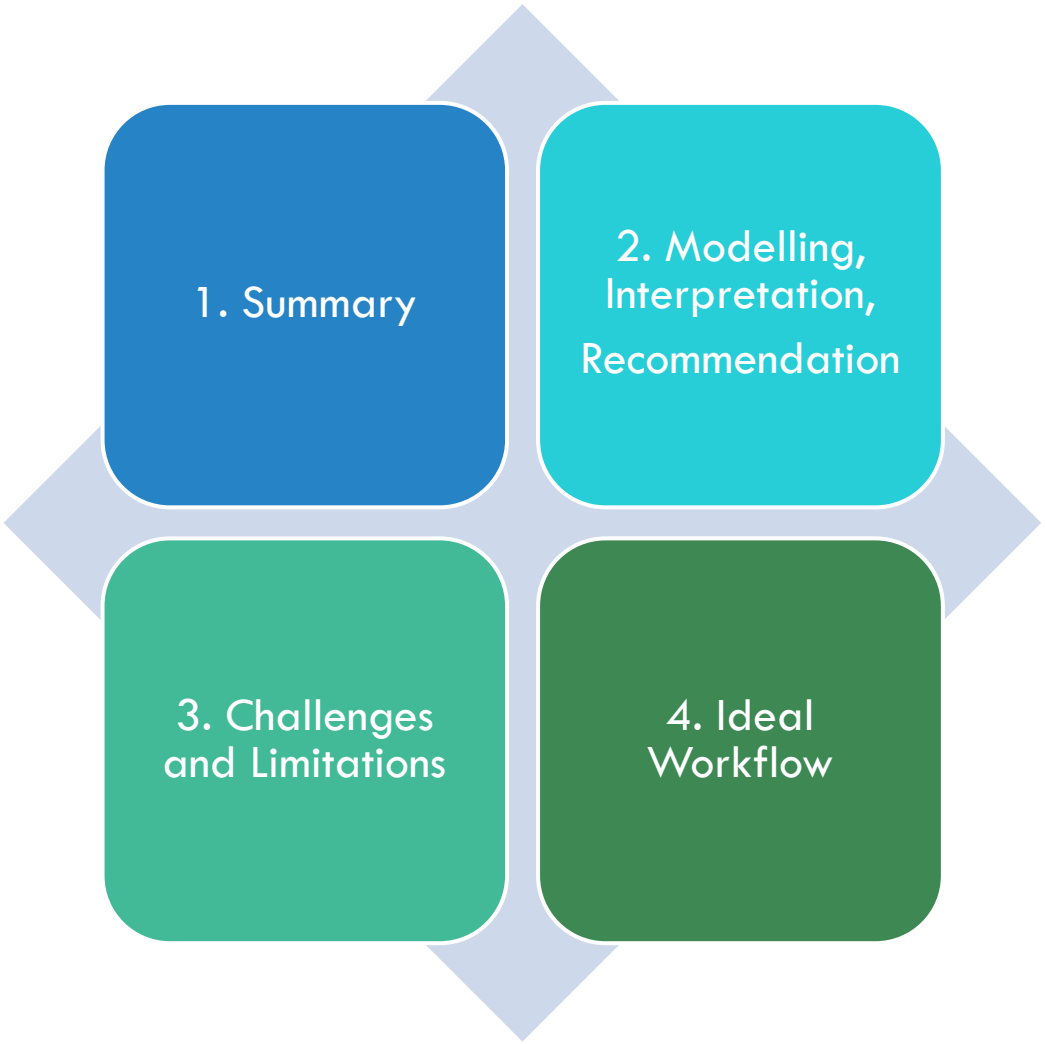
MOVIE INVESTMENT ANALYSIS

By Mengyue Zhao
beczhaozmy@gmail.com

BUSINESS CONTEXT

In this fictitious case, the owner of the City Movie Production company wants to produce a drama movie with director Pedro Almodóvar and translate it to English only. One condition for this investment is that the movie has to be produced in Spanish. What's more, the owner cannot afford lose money because the company is very small.

This analysis use machine learning techniques to create insight into this plan and challenge it.



PRESENTATION OVERVIEW

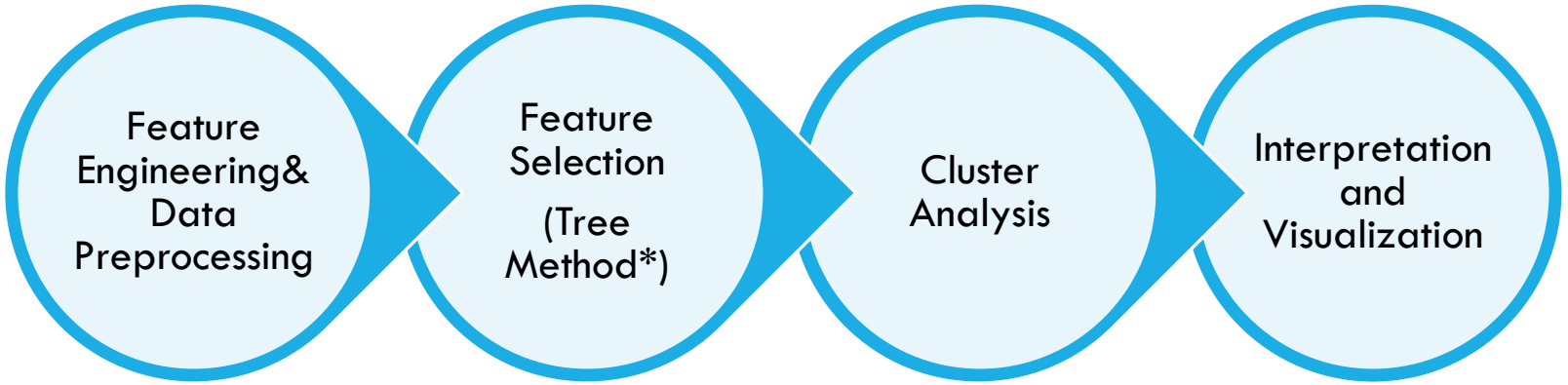
OUR MAIN GOAL:
1) LOW RISK
2) HIGH PROFIT

City Movie's Plan:

- Drama Film In Spanish, Translated to English Only

- Director Pedro Almodóvar

WORK FLOW OVERVIEW



FEATURE ENGINEERING

1

- Number of languages spoken, has additional language

2

- Retrieve Director Information and cast information from data in json format.

3

- get whether a movie have “Big Directors”; get whether a movie have “popular star”.

- These are inferred from 2)

MAIN CHALLENGES & ROADBLOCKS

Missing Values in the Target

- No revenue and budget data for Spanish Films in the recent decade. (87% shows 0 revenue)
- **Solution: Use popularity* as a proxy.**

Data Leakage

- Computed "Blockbuster Directors" from "total voting", which is colinear with the target.
- **Solution: Select the top directors that directed the most movies as "big directors" instead.**

ANALYTIC DATA VIEW

- Dataset: “Movie Data Set” from Kaggle, linking “credits” and “metadata”

- Number of observations: 660 Spanish Movie that are released between 2010-Jan and 2019-Dec

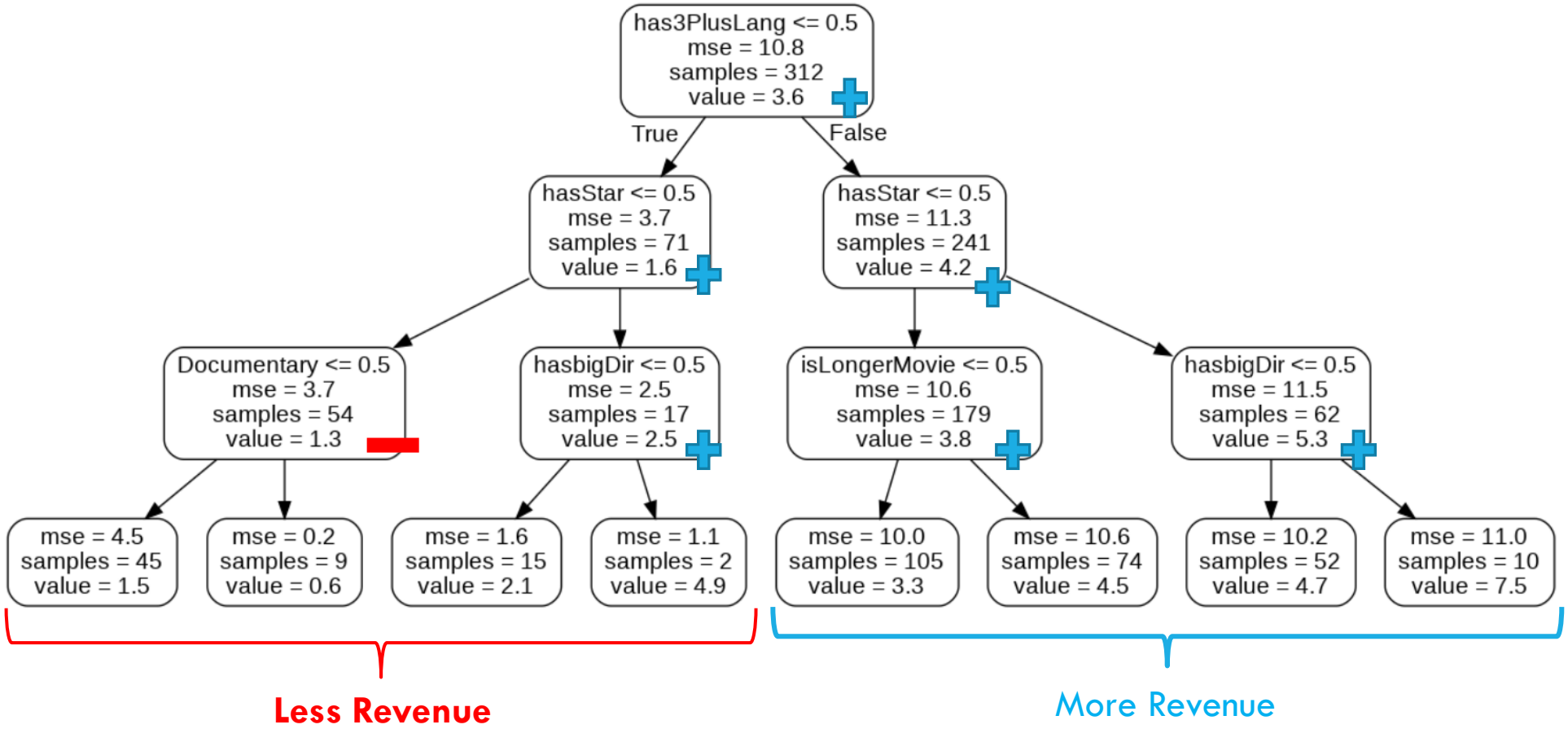
Target	Other Aspects		Crew and Cast		16 Genres			
popularity	has3PlusLang	isLongerMovie	hasStar	hasbigDir	Action	Adventure	Animation	…
0.134014	1	0	1	0	1	0	0	0
0.134014	1	0	1	0	0	1	0	0
2.66762	1	0	0	0	0	0	1	0
6.7106	1	1	1	0	0	0	0	1
11.0214	1	1	0	0	0	0	0	0

FEATURE IMPORTANCE

Remarks – These factors are important for the success of a movie:

- having more than three spoken languages
  - (more potential markets)
- has a star in its cast
- has longer duration
- Genres are less important in general but Comedy, Adventure, Thrillers, etc. do affect the box office.

```
Variable: has3PlusLang          Importance: 0.2
Variable: isLongerMovie         Importance: 0.17
Variable: hasStar               Importance: 0.11
Variable: hasbigDir             Importance: 0.1
Variable: Comedy                Importance: 0.09
Variable: Adventure             Importance: 0.05
Variable: Action                Importance: 0.04
Variable: Drama                 Importance: 0.04
Variable: Thriller              Importance: 0.04
Variable: Crime                 Importance: 0.03
Variable: Documentary           Importance: 0.03
Variable: Family                Importance: 0.02
Variable: Animation             Importance: 0.01
Variable: Fantasy               Importance: 0.01
Variable: History               Importance: 0.01
Variable: Music                 Importance: 0.01
Variable: Mystery               Importance: 0.01
Variable: Romance               Importance: 0.01
Variable: War                   Importance: 0.01
Variable: Horror                Importance: 0.0
Variable: None                  Importance: 0.0
Variable: Science Fiction       Importance: 0.0
Variable: TV Movie              Importance: 0.0
```

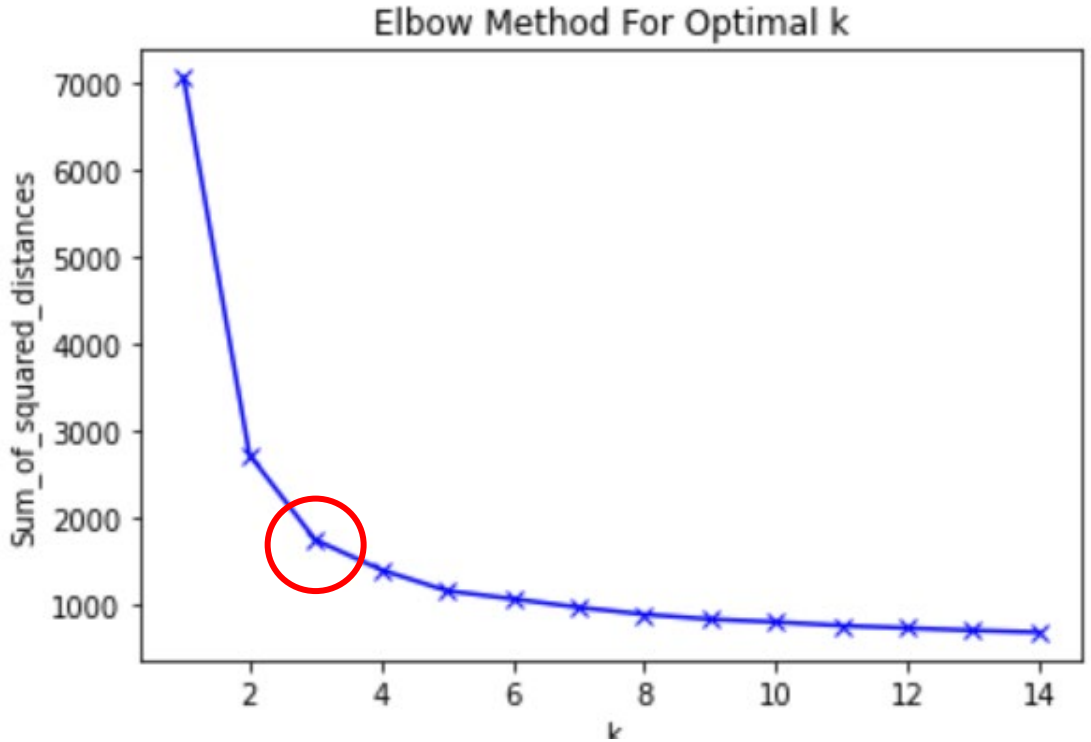


FEATURE IMPORTANCE

CLUSTER ANALYSIS: ADV

	has3PlusLang	isLongerMovie	hasStar	hasbigDir	Comedy	Action	Adventure	Drama	Crime	Documentary	popularity
0	1	0	1	0	0	0	0	1	0	0	0.134014
1	1	0	1	0	0	0	0	1	0	0	0.134014
2	1	0	0	0	0	0	0	0	0	1	2.66762
3	1	1	1	0	0	0	0	0	0	0	6.7106
4	1	1	0	0	0	0	0	1	0	0	11.0214

CLUSTER ANALYSIS: ELBOW METHOD



cluster Name	clusterLabel	popMinMax	has3PlusLang	isLongerMovie	hasStar	Comedy	hasbigDir	Adventure	Drama	Action	Documentary
"Avoid at All Cost"	0	0.314683	0.757576	0.026515	0.185606	0	0.075758	0.037879	0.481061	0.034091	0.109848
"BlockBuster"	1	0.747359	0.910995	0.931937	0.282723	0	0.125654	0.015707	0.73822	0.04712	0.015707
"Medium"	2	0.54667	0.746411	0.277512	0.253589	1	0.114833	0	0	0	0

CLUSTER ANALYSIS

RECOMMENDATION

Expected: "Big Directors" + Good Cast

- Also: drama are likely to do well.
- It is unclear how Pedro Almodóvar contributes to the box office, but big directors are much safer options if we want to avoid risk at all cost.

Unexpected Trends:

- Romance is not an important genre that contributed to popularity.
- Documentaries make important contributions to poor box office.
- People like Longer Movies.

LIMITATIONS

Lack of Information

- Box office information is never transparent.
- There is no budget allocation or revenue information.

- Potential Problem: Blockbuster Directors are obtained from top 1 percentile popularity. This might create a data leakage program since popularity is the target.

- Additional features: visual effect, Cost on scenes, advertisement budget, etc.

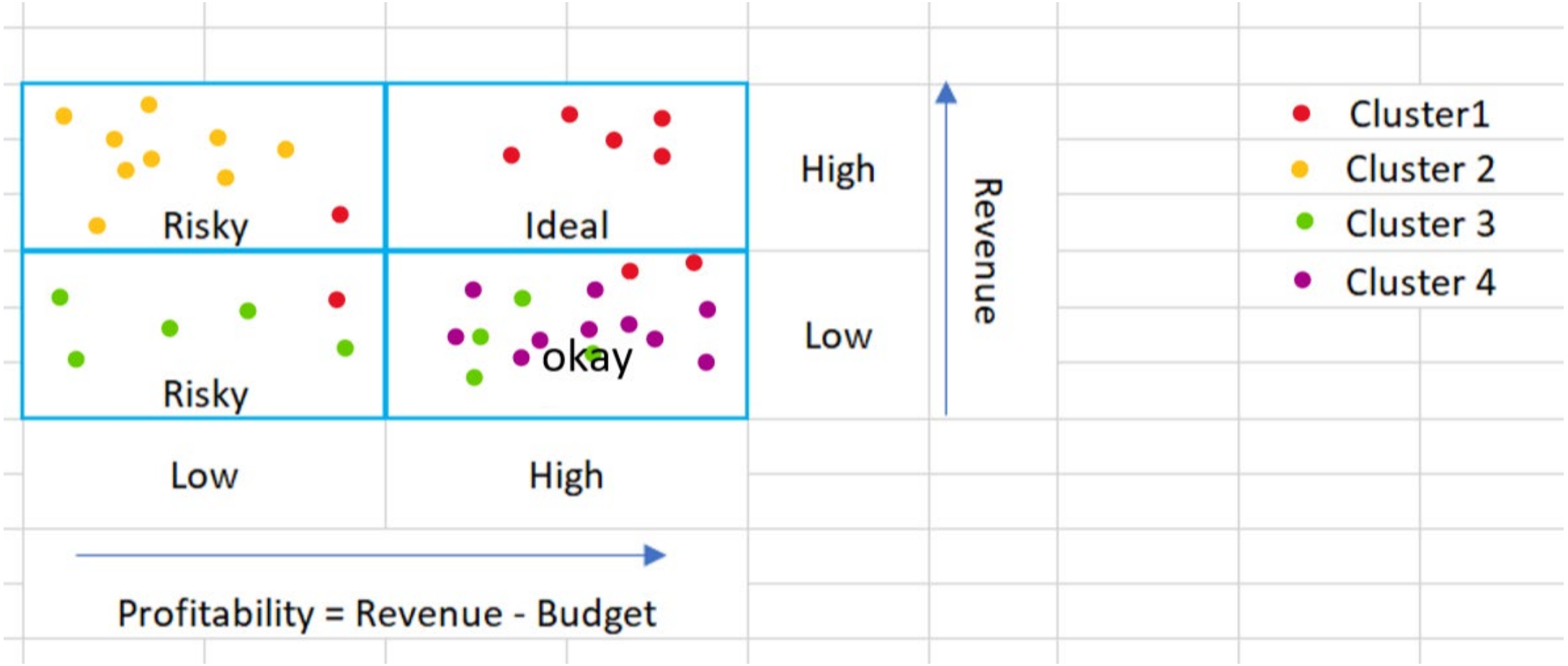
ADDITIONAL STEPS

- Textual Analysis on keywords, title, tagline
- Better Analysis on spoken language.

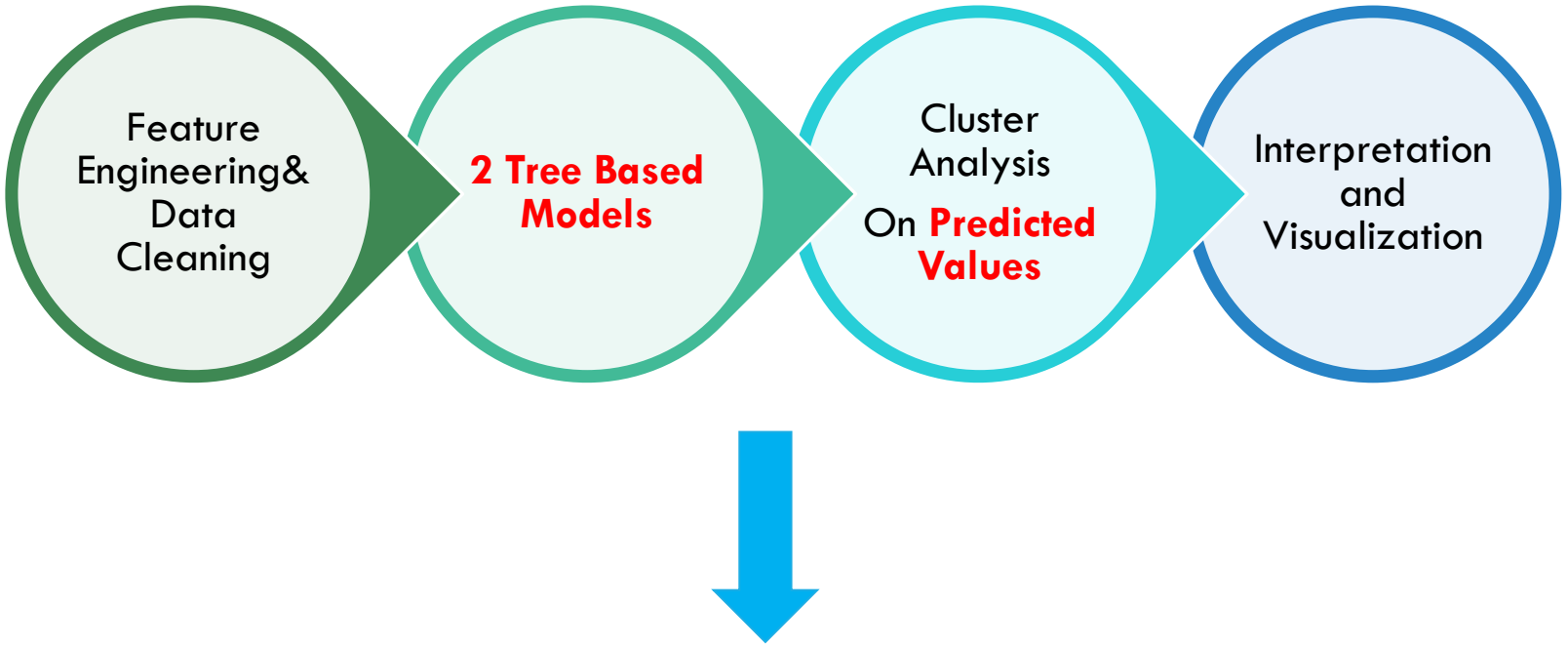
- Retrieve more information about visual effect, and other aspects of film production.

- Begin the presentation with an EDA on the Director, the Genre, and other outstanding features that interest the audience.

IDEAL INSIGHTS



IDEAL WORK FLOW



THANKS