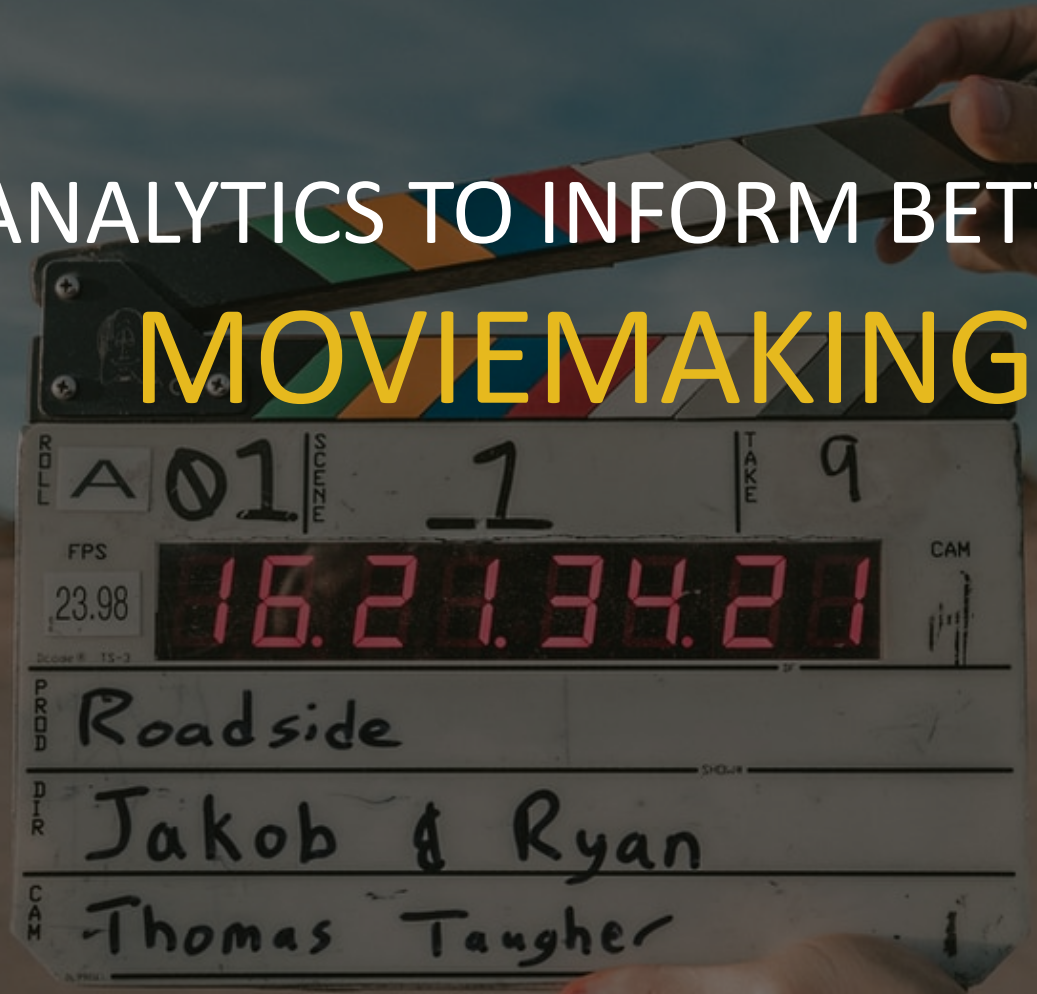
ANALYTICS TO INFORM BETTER MOVIEMAKING



TEAM 5

COMM 4559: BIG DATA & MARKETING ANALYTICS

Sterling Clay, Leanne Musa, Hannah Reeves, JJ Sharma, Yulin Yu

On our honor as students, we have neither given nor received aid on this assignment

MEET THE TEAM

Sterling Clay
Commerce ‘21

Leanne Musa
Commerce ‘21

Hannah Reeves
Commerce ‘21

JJ Sharma
Commerce ‘21

Yulin Yu
Statistics ‘21

ROADMAP

01

INTRODUCTION

Learn more about us and the business problems to be addressed during the presentation

02

DATASET

Dive into the data that drives our statistical models and influences decision-making

03

DESCRIPTIVE ANALYTICS

Discover the attributes that make up the top 100 English-speaking movies based on over 120,000 reviews

04

PREDICTIVE ANALYTICS

Harvest the power of predictive modelling using machine learning to inform sentiment based on review

05

CONCLUSION

See how these analytical methodologies can be used by filmmakers, streaming platforms and more

CONTEXT OF THE BUSINESS PROBLEM

We want to uncover the most significant movie features and characteristics that influence a movie to become top-rated. Through our analysis, we intend to provide answers to the following business questions:

What are the key attributes of the top 100 movies, and how can they be reproduced?

How do movie producers select which movie scripts to go through with?

Why is this business problem important, and what are the implications?


DATASET

COLLECTION AND OVERVIEW

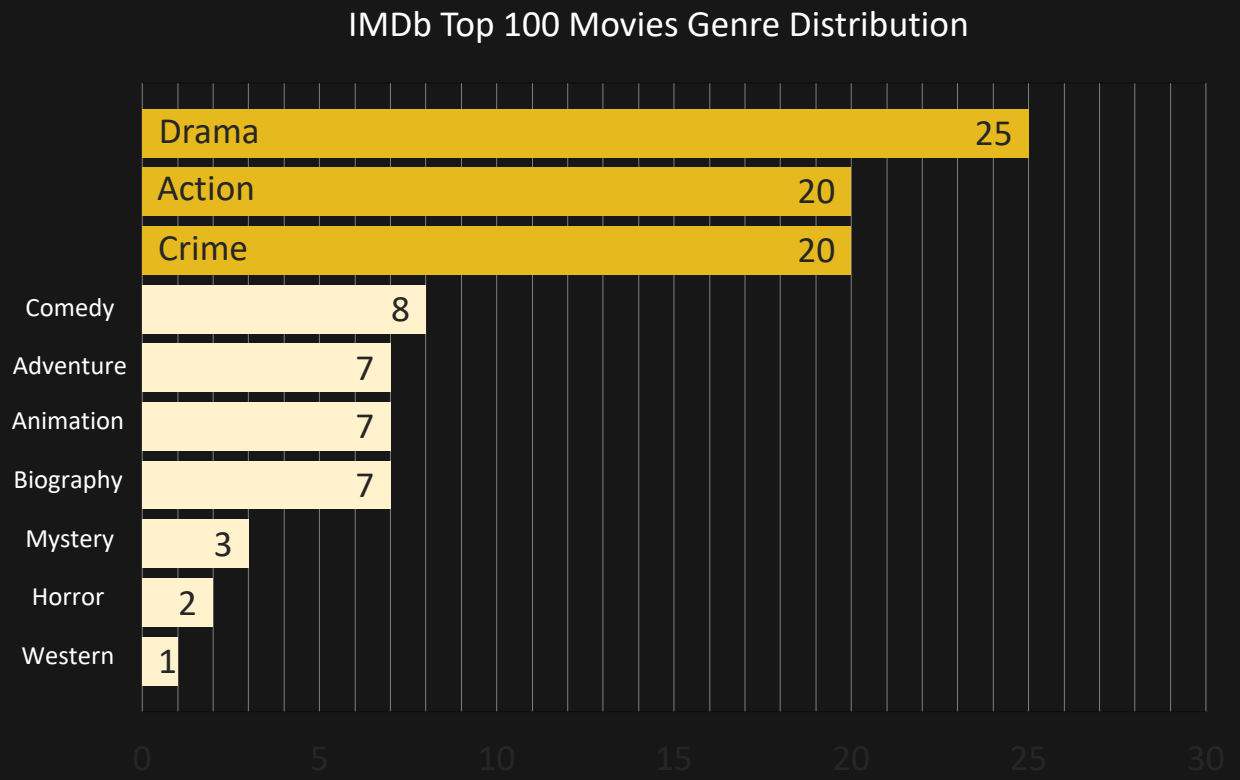
127K

Individual Reviews

ATTRIBUTES

- Movie ID
- Title
- Year
- Length
- Genre
- Rating
- Gross Income

MOVIE AND GENRE DATA



2H15M

Mean Movie Length

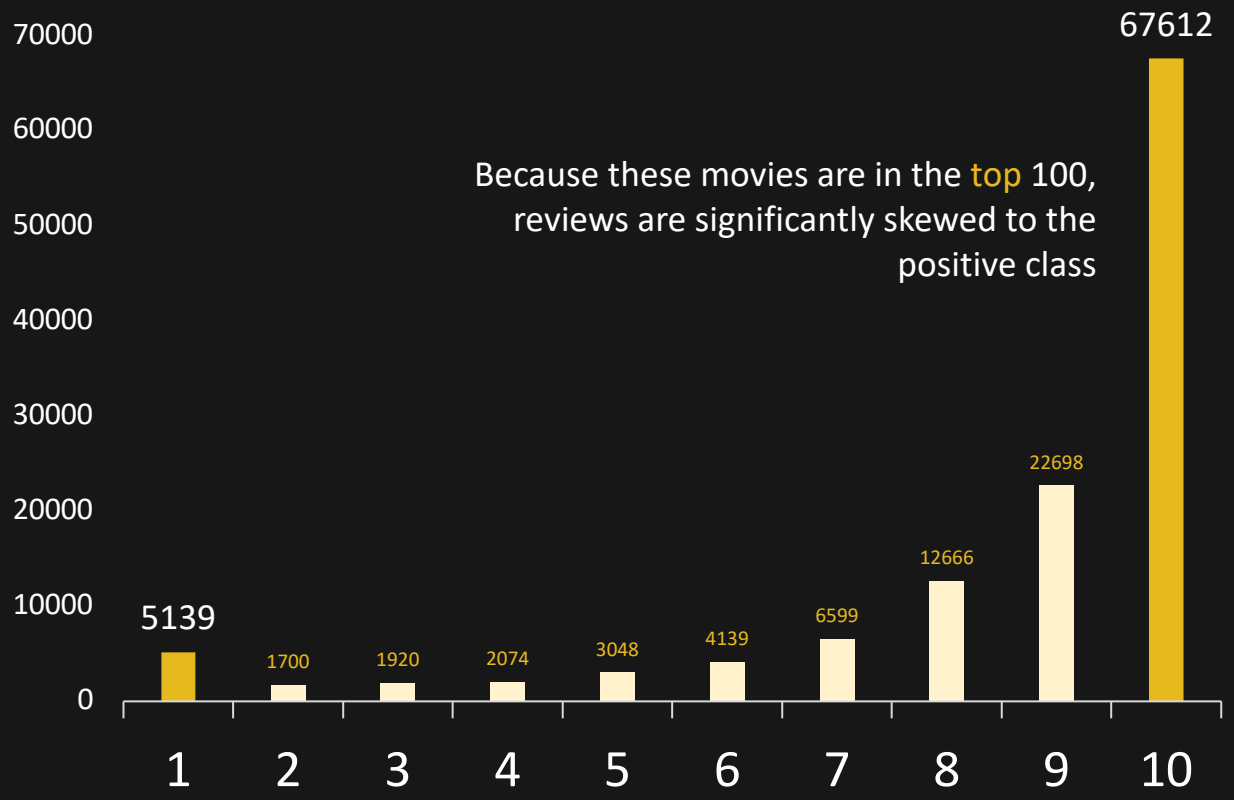
136.88 MIL

Mean Gross Income

REVIEW DATA

127,649

Number of Reviews

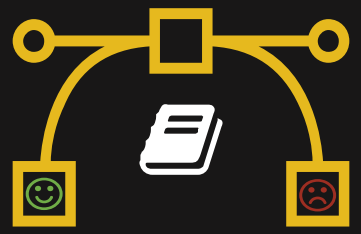


Because these movies are in the top 100, reviews are significantly skewed to the positive class


DESCRIPTIVE
ANALYTICS

EMOTION ANALYSIS

NRCLexicon



MIT-approved Python software project used to predict the emotion and sentiment (positive or negative) of a given text.

27K

Word Count incl associated emotions

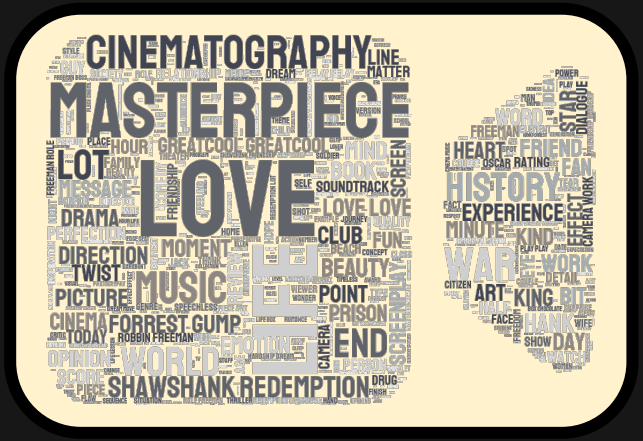
EMOTIONS

Anger
Fear
Anticipation
Trust
Surprise
Sadness
Joy
Disgust

MOST REPRESENTED EMOTIONS IN TOP 3 GENRES

DRAMA

JOY



ACTION

ANTICIPATION

CRIME

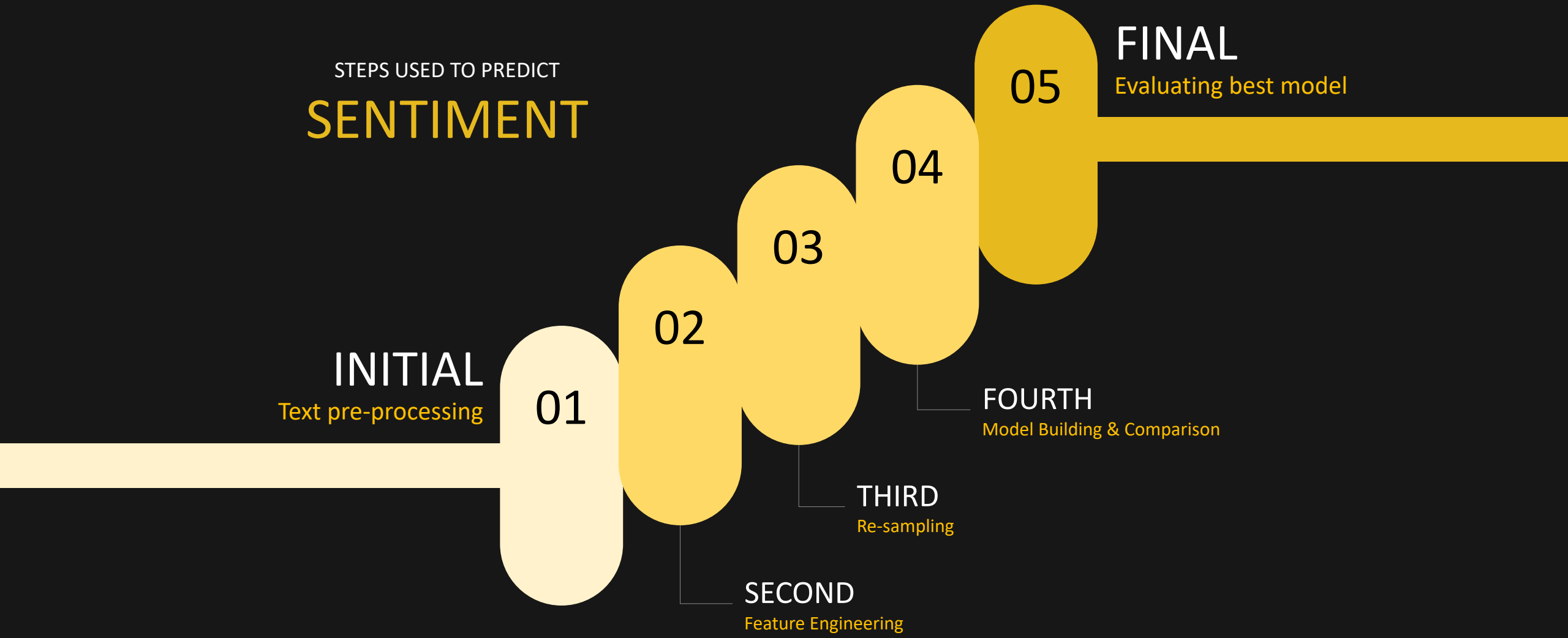
FEAR


PREDICTIVE
ANALYTICS

PREDICTIVE ANALYTICS

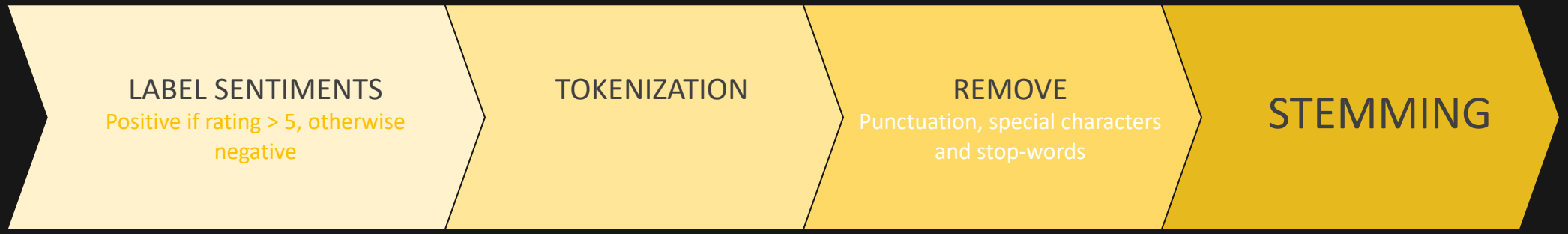
STEPS USED TO PREDICT

SENTIMENT



SENTIMENT ANALYSIS PROCESS

STEP 01 | TEXT PRE-PROCESSING



SENTIMENT ANALYSIS PROCESS

STEP 02 | FEATURE ENGINEERING

BAG OF WORDS

Representation of text

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

TF-IDF

Term Frequency-Inverse Document Frequency

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

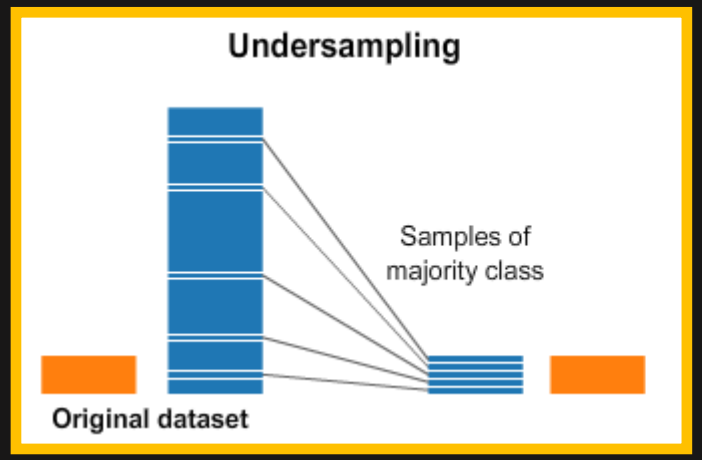
SENTIMENT ANALYSIS PROCESS

STEP 03 | RESAMPLING

UNDERSAMPLING

Near Miss

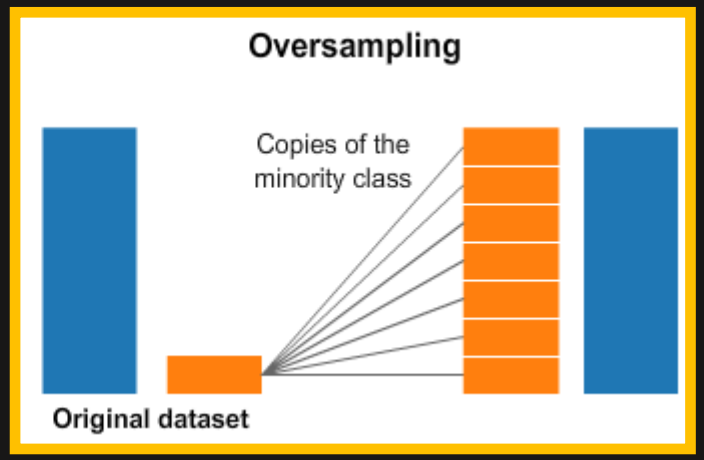
Remove the instances of the majority class when instances of two different classes are very close to each other



OVERSAMPLING

Synthetic Minority Over-Sampling Technique

Randomly select one or more of the k-nearest neighbors for each example in the minority class



SENTIMENT ANALYSIS PROCESS

STEP 04 | MODEL BUILDING & COMPARISON

20

Classification Models

MODELS USED

Logistic Regression
Support Vector Machine
Multinomial Naïve Bayes
XGBoost
Decision Tree

Model	Precision	Recall	F-1 Score	Accuracy
MNB_TFIDF_Over	**0.96**	**0.96**	**0.96**	**0.96**
MNB_BOW_Under	0.87	0.86	0.86	0.86
XGB_BOW_Over	**0.95**	**0.95**	**0.95**	**0.95**
XGB_BOW_Under	0.83	0.83	0.83	0.83
Logistic_Regression_TFIDF_Over	**0.95**	**0.94**	**0.94**	**0.94**
Logistic_Regression_TFIDF_Under	0.87	0.87	0.87	0.87
SVM_BOW_Over	**0.94**	**0.94**	**0.94**	**0.94**
SVM_BOW_Under	0.87	0.87	0.87	0.87
Decision Tree_TFIDF_Over	**0.89**	**0.89**	**0.89**	**0.89**
Decision Tree_BOW_Under	0.71	0.71	0.71	0.71

MNB_TFIDF_OVER

Best Performing Model

MULTINOMIAL NAÏVE BAYES

STEP 05 | EVALUATING BEST MODEL

Given the dependent feature vector (x_1, …, xn) and the class Ck. Bayes' theorem is stated mathematically as the following relationship:

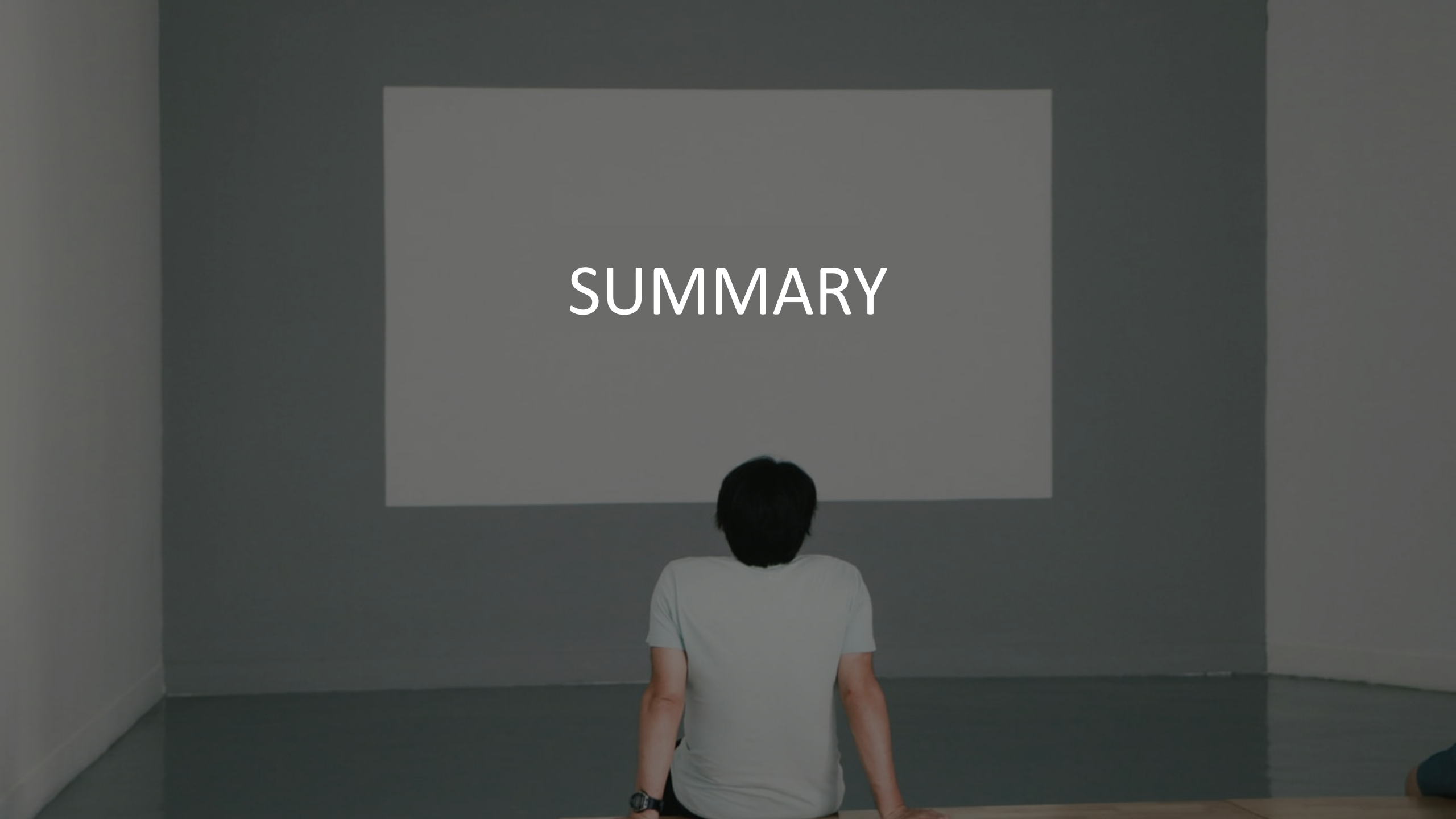
$$P(C_k \mid x_1, \ldots, x_n) = \frac{P(C_k)P(x_1, \ldots, x_n \mid C_k)}{P(x_1, \ldots, x_n)}$$

Posterior probability = (class prior probability x likelihood) / predictor class prior probability

Advantages

It is simple and easy to implement.

It is fast and can be used to make **real-time predictions**.


SUMMARY

LIMITATIONS

Reviews were solely from IMDB

Dataset only included movies in English

Focus was only on “top movies”

Our sentiment analysis only gives basic insight on cast and directors

DATA INSIGHTS

- The descriptive statistics suggest that the common features of the top 100 English Movies are:
  - Genres (drama, crime, action)
  - Length (100-120 min)
- Emotion analysis informs us about the audience's reaction to movie scenes
- Our machine learning models yield a high accurate prediction on whether a movie review is positive/negative
  - Oversampled data performed better than the undersampled data in general
  - Simple models are sufficient to get high accuracy; also easy to implement and interpret.

FUTURE APPLICATIONS

Personalized movie recommendations on streaming platforms based on audience reviews

Help directors and producers make more informed decisions about song selection, sponsorships, and other product placements


QUESTIONS?
THANK YOU!