Movie Schedule Optimization
Using Sentimental Analysis of Movie Reviews

Group 15
Cao, Jiali A0232321Y
Huang, Hai-Hsin A0231906J
Tsao, Kai-Ting A0231947Y
Mingxuan Yang A0231854E

Agenda

- **Background**
- **Problem Statement**
- **Methodology**
  - OLS Linear Regression
  - Sentimental Analysis Using TextBlob
  - Predict Movie Sentiments Using Classification models
- **Business Application**
- **Conclusion and Further Navigation**
- **Reference**

Background

Motivation

Traditionally, theaters would predict a high box office for a movie with famous producers, popular cast and high budget

With the spread of SNS, people tend to search for comments on the Internet before buying a ticket.

People's comments become more and more important to affect the revenue of a movie.

Problem Statement

- Does the ranking on movie rating websites affect potential audiences' willingness to watch the movie?
- How to find a model to analyze audiences' sentiment and help to make film arrangement?

Methodology

Linear Regression

to evaluate the effects of Internet rating on movie revenue

Sentimental Analysis Using TextBlob

to explore the polarity and subjectivity of movie review contents

Predict Movie Sentiments Using Classification models

Compare Logistic Regression, SVM, Naive Bayes and K-nearest Neighbors Classifier

Methodology

Linear Regression

to evaluate the effects of Internet rating on movie revenue

Sentimental Analysis Using TextBlob

to explore the polarity and subjectivity of movie review contents

Predict Movie Sentiments Using Classification models

Compare Logistic Regression, SVM, Naive Bayes and K-nearest Neighbors Classifier

Linear Regression

Data Source	TMDB and GroupLens
Time Span	2000-2017
Data Size	2561
Dependent variable	revenue
Independent variable	vote_average
Control variables	budget, genre, year, runtime and country

Linear Regression

- R Squared: 0.501
- Coef: 0.4747
- P Value: 0

• vote_average has a significant positive correlation with movies' revenue. The result matches our hypothesis

Methodology

Linear Regression

to evaluate the effects of Internet rating on movie revenue

Sentimental Analysis Using TextBlob

to explore the polarity and subjectivity of movie review contents

Predict Movie Sentiments Using Classification models

Compare Logistic Regression, SVM, Naive Bayes and K-nearest Neighbors Classifier

Sentimental Analysis

Datasets

Data source: Rotten tomatoes review
Time span: 2011 - 2020 (10 years range)
Data size: 59,498 rows , 8 columns

Movie ID	critic_name	top_critic	publisher_name	review_type	review_score	review_date
m/0814255	Greg Maki	FALSE	Star-Democrat (Easton, MD)	Rotten	D+	2011/11/5
m/100001312	Dennis	TRUE	Dennis Movie Reviews	Fresh	B	2011/5/12

review_content
The premise of Percy Jackson & the Olympians: The Lightning Thief holds great potential. Potential the film never realizes.
Lumet keeps things tense, sweaty, suspenseful and entertaining despite the contrived story line.

Sentimental Analysis

Feature Engineering

Encode review_rank to review_score

review_rank	review_score
A	12
B+	11
B	10
B–	9
…	…
F	1

Encode review_type to sentiment

review_type	sentiment
Fresh →	1
Rotten →	0

Fresh: review score >= 9
Rotten: review score <9

Sentimental Analysis

Polarity and Subjectivity

We used a package called TextBlob to analyze the polarity and subjectivity of each review content.

Polarity

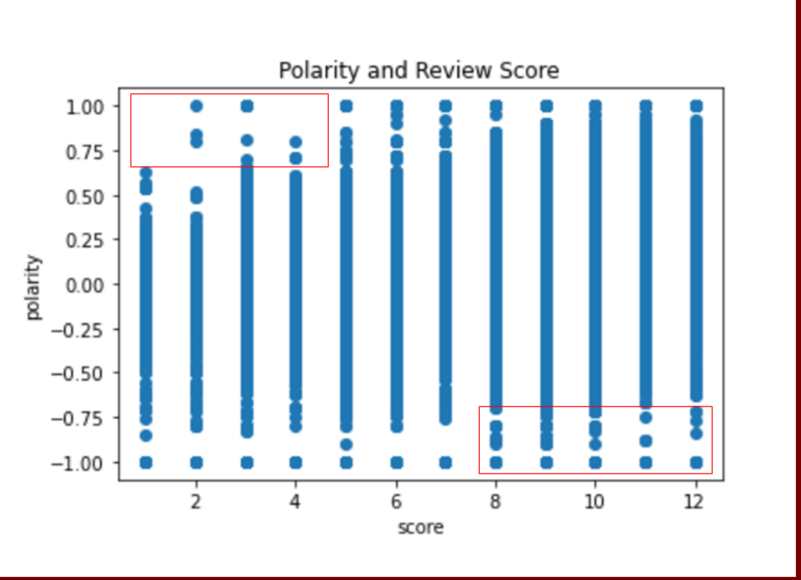
-1 Negative **0 Neutral** **+1 Positive**

Subjectivity

0 Objective **+1 Subjective**

Sentimental Analysis

EDA and Preliminary Results

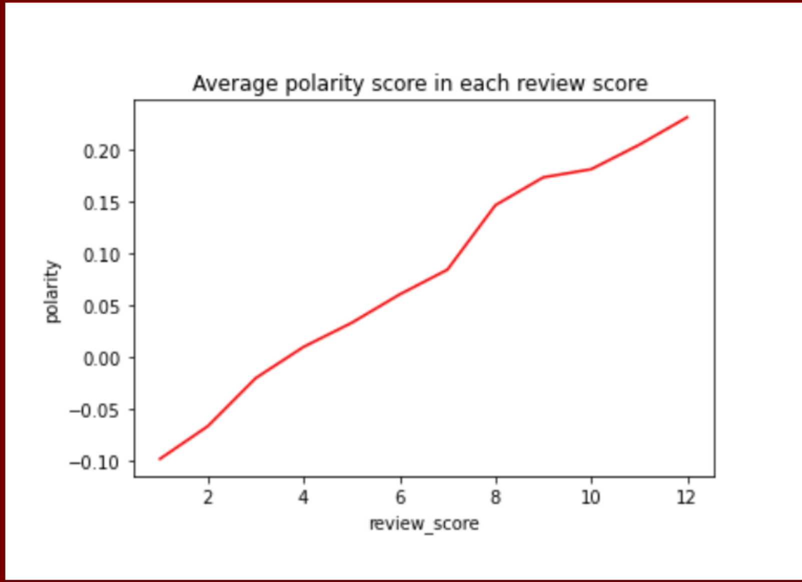


Polarity score distribution within each review score

More data points with a positive polarity score in higher review scores and less data points with positive polarity scores in lower review scores.

Sentimental Analysis

EDA and Preliminary Results



Positive correlation between the polarity score and review score

We computed the average polarity score in each review score range. Interestingly, the higher the review score, the higher the polarity score.

Sentimental Analysis

EDA and Preliminary Results

Model:	OLS	Adj. R-squared:	0.711
Dependent Variable:	score	AIC:	205112.6471
Date:	2021-11-17 18:48	BIC:	205148.6219
No. Observations:	59498	Log-Likelihood:	-1.0255e+05
Df Model:	3	F-statistic:	4.869e+04
Df Residuals:	59494	Prob (F-statistic):	0.00
R-squared:	0.711	Scale:	1.8394

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Intercept	5.1997	0.0100	521.0694	0.0000	5.1801	5.2192
polarity	0.6180	0.0209	29.5853	0.0000	0.5771	0.6589
sentiment	4.3658	0.0121	361.6722	0.0000	4.3421	4.3895
top_critic_dummy	0.0038	0.0137	0.2780	0.7810	-0.0231	0.0307

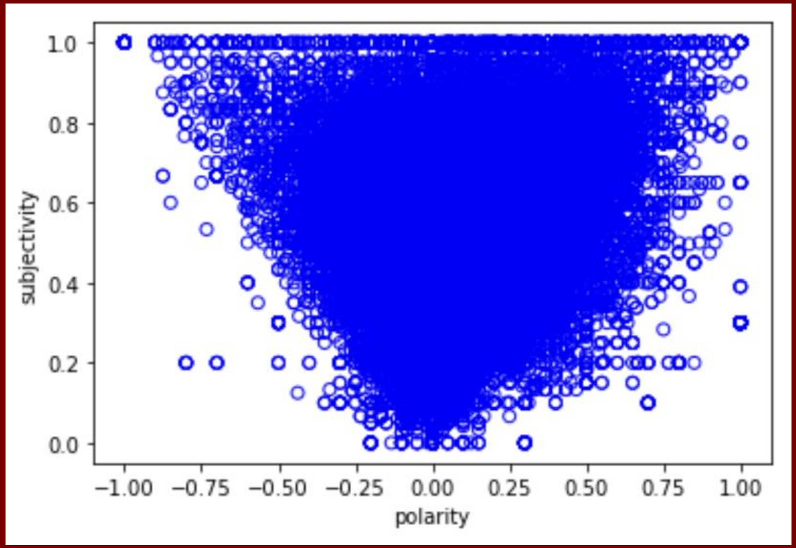
Omnibus:	749.086	Durbin-Watson:	1.617
Prob(Omnibus):	0.000	Jarque-Bera (JB):	740.345
Skew:	-0.252	Prob(JB):	0.000
Kurtosis:	2.789	Condition No.:	5

The polarity score has a significant effect on the review score

The polarity score has a positive coefficient and a 0.0000 p-value, meaning that the polarity score has a significant effect on the review score, and they are positively correlated.

Sentimental Analysis

EDA and Preliminary Results

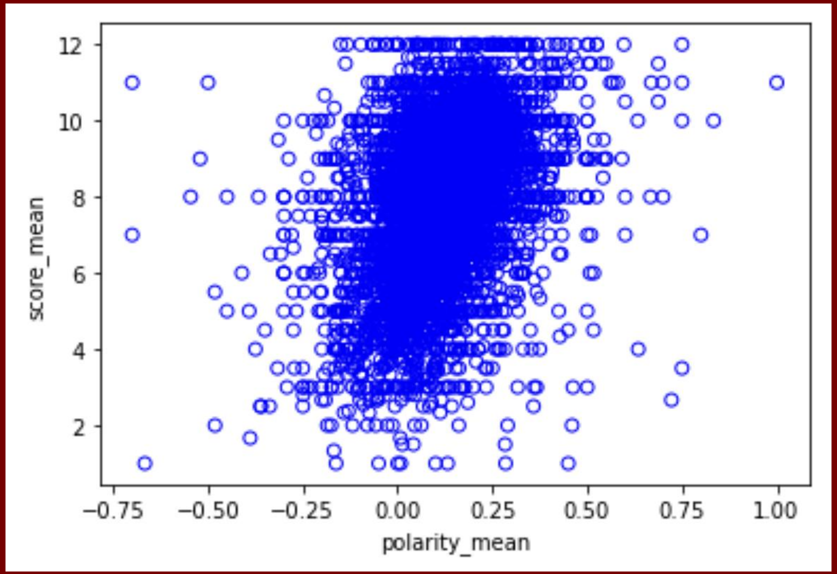


Relationship between polarity and subjectivity

More polar comments tend to be more subjective.

Sentimental Analysis

EDA and Preliminary Results



Relationship between average review score for each movie and average polarity score for each movie

Comments that are too subjective (subjectivity score > 0.8) and movies that have only one comment are filtered out. There is a slightly positive correlation.

Methodology

Linear Regression

to evaluate the effects of Internet rating on movie revenue

Sentimental Analysis Using TextBlob

to explore the polarity and subjectivity of movie review contents

Predict Movie Sentiments Using Classification models

Compare Logistic Regression, SVM, Naive Bayes and K-nearest Neighbors Classifier

Classification models

Introduction

Predict Movie Sentiments
With Review Contents
Using Classification models

Using classification models, including Logistic Regression, SVM, Naive Bayes and K-nearest Neighbors Classifier to train the same dataset (Rotten tomatoes review)

Review Contents → **Movie Sentiments**

Classification models

The count vectorizer & the tf-idf vectorizer

How can we input review contents as independent variables?

COUNT VECTORIZER

The count vectorizer considers the frequencies of words in a sentence.

TF-IDF VECTORIZER

The tf-idf vectorizer considers both the frequencies a word appears in a sentence and the number of sentences the word appears in.

Classification models

Logistic Regression

SVM

Multinomial Naive Bayes

K-nearest Neighbors Classifier

Classification models

Accuracy rate comparisons

Accuracy	Count vectorizer for bag of words (BOW)		Tfidf vectorizer	
	Test accuracy rate	Training accuracy rate	Test accuracy rate	Training accuracy rate
Logistic Regression	0.6479	0.9322	0.6458	0.6576
SVM	0.6458	0.8618	0.6458	0.6577
Naïve Bayes (Multinomial NB)	0.6568	0.9338	0.6473	0.9338
K-nearest neighbors classifier	0.6461	0.6577	0.6458	0.6577

Classification models

Accuracy rate comparisons

Accuracy	Count vectorizer for bag of words (BOW)		Tfidf vectorizer	
	Test accuracy rate	Training accuracy rate	Test accuracy rate	Training accuracy rate
Logistic Regression	0.6479	0.9322	0.6458	0.6576
SVM	0.6458	0.8618	0.6458	0.6577
Naïve Bayes (Multinomial NB)	0.6568	0.9338	0.6473	0.9338
K-nearest neighbors classifier	0.6461	0.6577	0.6458	0.6577

Classification models

Accuracy rate comparisons

Accuracy	Count vectorizer for bag of words (BOW)		Tfidf vectorizer	
	Test accuracy rate	Training accuracy rate	Test accuracy rate	Training accuracy rate
Logistic Regression	0.6479	0.9322	0.6458	0.6576
SVM	0.6458	0.8618	0.6458	0.6577
Naïve Bayes (Multinomial NB)	0.6568	0.9338	0.6473	0.9338
K-nearest neighbors classifier	0.6461	0.6577	0.6458	0.6577

Classification models

Unbalanced data

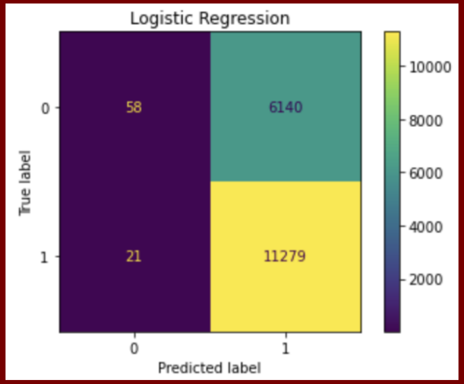
Sentiment =1
(FRESH)

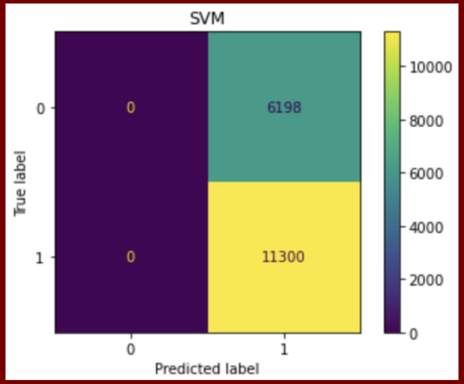
64.58 %

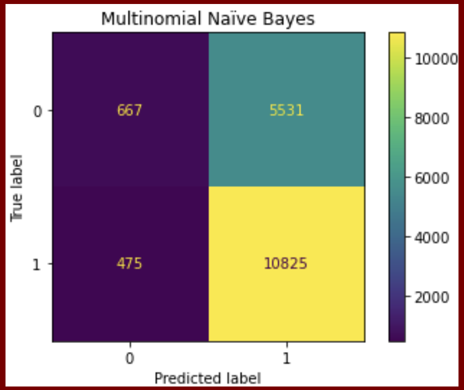
Sentiment = 0
(ROTTEN)

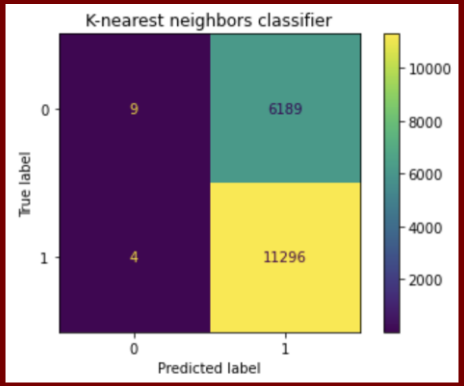
Classification models

Confusion Matrix



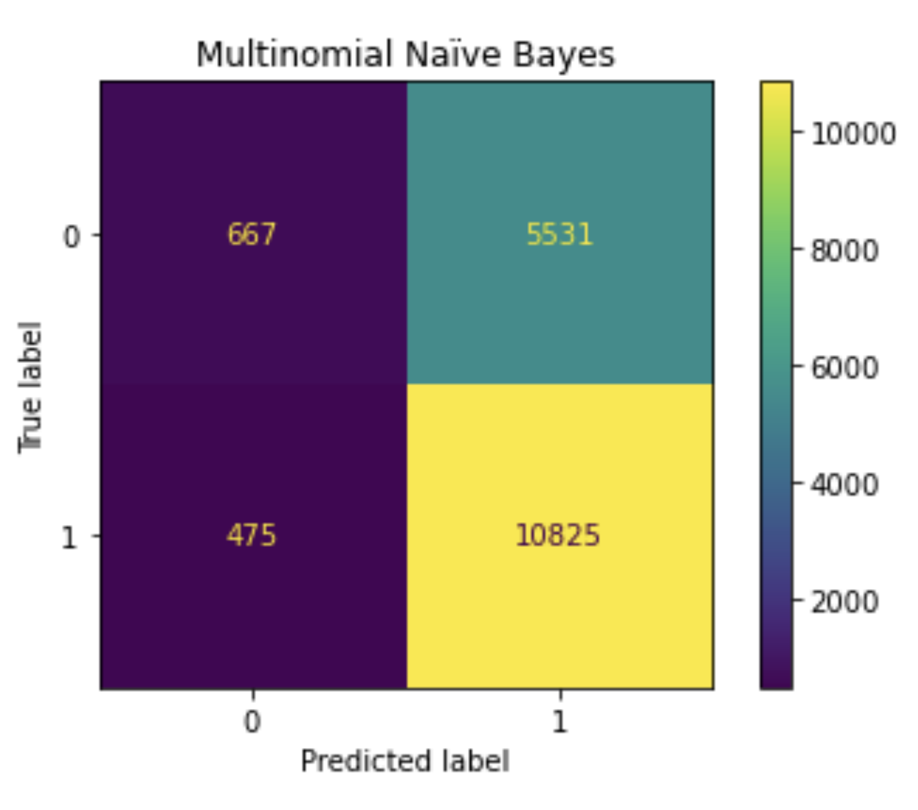




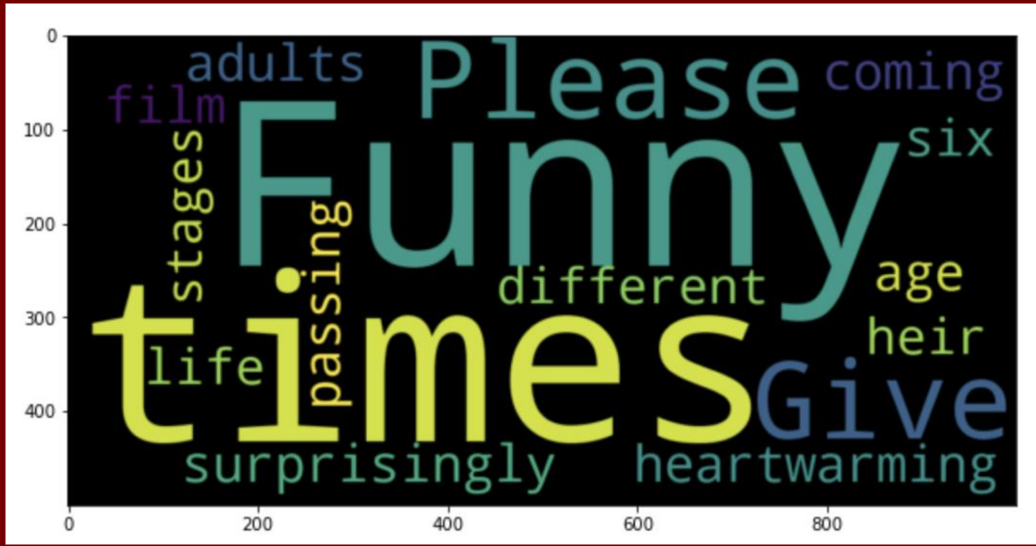


Confusion Matrix

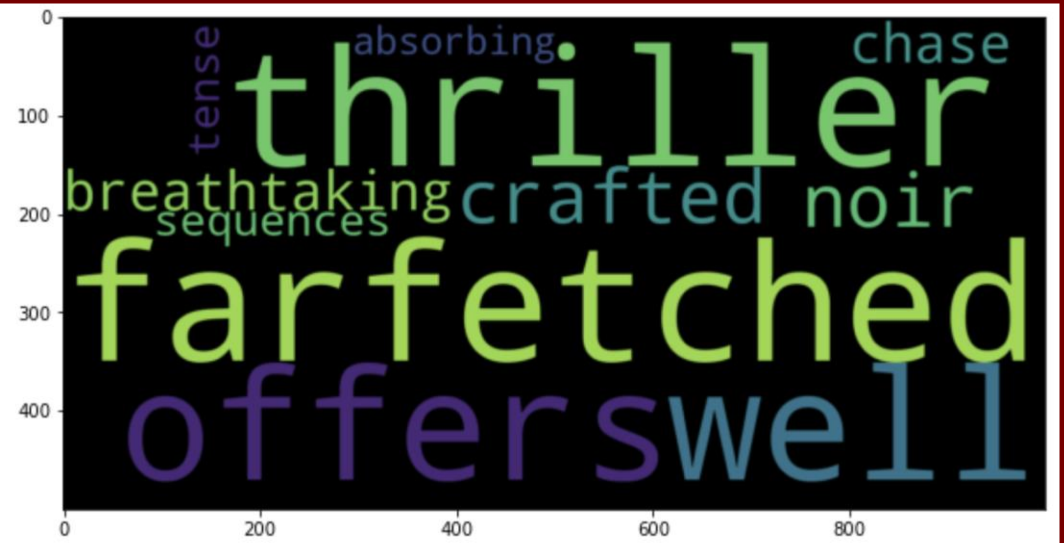
Multinomial NB



Word Cloud

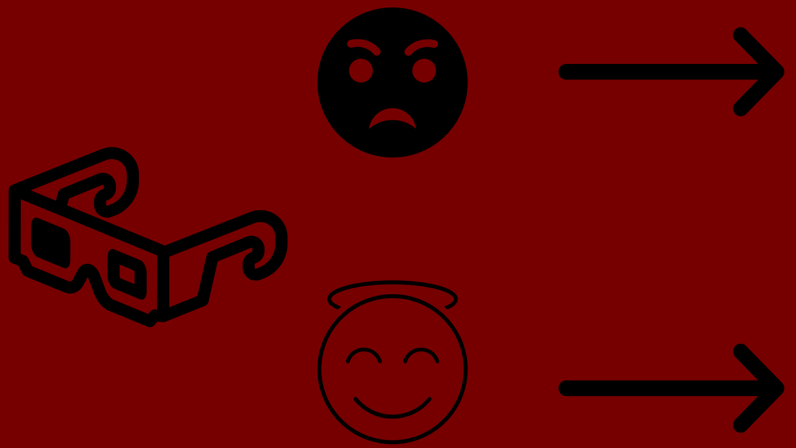


Positive review words



Negative review words

Business Application



Further Editing

Target Demographics:
- Locations
- Airtimes

Conclusion and Further Navigation

"Too subjective"
And
"Too polarized"?

Does tokenized words really mean what the commentor wants to imply?

Are sentiments simply "positive" or "negative"?

Reference

[1] Rotten Tomatoes movies and critic reviews dataset

[2] The Movies Dataset

[3] Python – Text Classification using Bag-of-words Model

[4] TF-IDF Vectorizer scikit-learn

Movie Schedule Optimization
Using Sentimental Analysis of Movie Reviews

Group 15
Cao, Jiali A0232321Y
Huang, Hai-Hsin A0231906J
Tsao, Kai-Ting A0231947Y
Mingxuan Yang A0231854E