Erasmus
School of
Economics

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis - Data Science and Marketing Analytics Program

Consumer Reception Analysis of Chinese Films in Western Markets: Based on Machine Learning Models

Name student	Yuewen Bai
Student ID	609756
Supervisor	Michel van de Velden
Second assessor	Bas Donkers
Date of final version	April 30, 2023

The content of this thesis is the sole responsibility of the author and does not reflect the view of either the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

Between 2006 and 2021, seven foreign-language films made it to the top 25 most popular movies in North America on Boxofficemojo.com, none of which were Chinese. This suggests that there is a cultural discount that affects the popularity of Chinese films among western audiences. To explore methods to reduce the cultural discount and enhance the acceptance of Chinese films in western markets, this study collects data from Douban.com and IMDb.com using web scraping techniques and employs machine learning models to analyze the impact of cultural differences and film quality on local reception. The findings reveal that Douban rating scores and the drama genre significantly influence local reception. Specifically, Chinese films with high Douban ratings are more likely to experience cultural discount, while drama films are more likely to be accepted by western audiences.

Keywords: data mining, machine learning, cultural discount, consumer reception

Table of Contents

CHAPTER 1. Introduction

1.1 Research background and motivation

According to the *2021 Theme Report* [1] from Motion Picture Association(MPA), the top three box office markets in 2021 were China ($7.3 billion, including online ticketing fees), North America ($4.5 billion), and Japan ($1.5 billion). Besides the success of the Chinese film market, Chinese films were once highly anticipated in western markets, especially the box office success of *Crouching Tiger, Hidden Dragon* (2000), which grossed $128 million in North America – more than any foreign-language film to date. This success was followed by other hits like *Hero* (2002, released in the U.S. in 2004), *Jet Li's Fearless* (2006), *Kung Fu Hustle* (2005), and *Iron Monkey* (1993, released in the U.S. in 2001), all still ranked among the top 25 foreign-language films marketed in North America (Boxofficemojo.com) [2]. All the success brings the potential to the Chinese film industry to challenge America's dominant position in global film markets.

The Chinese film industry is currently encountering a significant challenge in terms of gaining more traction in western markets. Between 2006 and 2021, seven new films, from India, Mexico and Korea, managed to secure spots in the top 25 popular foreign-language film list. This is a notable statistic considering that China is the world's third-largest producer of films, having produced a total of 2479 films during this same time period. In contrast, India produced 1753 films, Mexico produced 383 films, and South Korea produced 1167 films (the-numbers.com) [3], with each country having three and one new films respectively on the top 25 popular foreign-language film list (Rosen, 2021).

Despite having one Korean film in the list of the top 25 popular foreign-language films, the Korean film industry still struggles to gain popularity in western markets, similar to Chinese films. In contrast, Mexican and Indian films, being geographically closer to North America or sharing language similarities respectively, have higher popularity in these markets. There are two reasons behind this phenomenon: first, Chinese films' quality needs to be improved; second, Chinese films are not accepted in western markets, even with good quality. This is because, when media products are transported across different cultural markets, audiences

[1] MPA 2021 Theme Report
[2] Foreign Language Films in North America 2022 - Boxofficemojo.com (2022)
[3] the-numbers.com (2022)

tend to prefer media productions that come from familiar backgrounds and are easy to understand. Media studies have shown that media products moving across cultural boundaries are often subject to ***local reception*** processes (Ang, 1989; Liebes & Katz, 1986). Besides, the content value often lessens due to cultural distance, which is a ***cultural discount*** phenomenon (Hoskins & Mirus, 1988). Several studies indicate that the degree of cultural distance between the exporting and importing markets affects the extent of this devaluation (Cantor & Cantor, 1986; Fu & Sim, 2010). Cultural distance can be reduced through factors such as geographical proximity and language similarity (Craig, Greene & Douglas, 2005). Hence, improving local reception in foreign markets and reducing content value loss are essential to thriving Chinese films in western markets.

As Fu and Sim (2010) demonstrated, a thriving media industry attracts foreign talent and investment, offering greater opportunities for career growth and financial returns, and the continuous arrival of foreign inputs, such as actors, directors, scriptwriters, and set designers, brings in new ideas and skills, revitalizing the industry. For example, Hollywood's economic success is evident in the substantial growth of its coffers. U.S. distributor revenues from theatrical movie releases have increased more than threefold as a proportion of the total economic activity in the United States, as measured by GDP, from approximately $1.1 billion in 1970 to $39.8 billion in 2003 (Waterman, 2009). The benefits of expanding the overseas market for Chinese films can be substantial.

In addition, cross-culture predictability can be very important to movie producers. ***Cross-culture performance predictability*** refers to the extent to which the performance of media products in a culture can be predicted by the performance of the same products in another culture (Lee, 2006) It is partly because movie projects can be highly risky (De Vany, 2003) and, from a business perspective, higher levels of predictability mean lower levels of risks (Lee, 2008).

1.2 Research questions

In this study, it is defined that, if the cultural discount is performed on a film, the film is not accepted by the foreign markets; otherwise, if a film shows equal or higher value than in the original market, which is defined as cultural premium, the film is accepted.

Despite being the world's third-largest film producer, having made 2479 new films between 2006 and 2021, the Chinese film industry is struggling to gain more popularity in western markets. In this study, the focus is on the relationship between local reception and cultural economics, with the main research question being: ***How to improve western consumers' reception of Chinese films?***

Audiences' reception of a film is decided by various aspects. Variances in cultural values, language, aesthetic preferences, and other factors can cause diverse perceptions of media products' quality in a foreign market. For example, foreign audiences may not be familiar with the domestic film cast, which is likely to decrease the star effect impact on audience reception; also, foreign audiences may be only interested in certain types of films, so some types of films may not be accepted in a foreign market. However, the values of media products in a foreign market are not completely predictable by their performance in their original market (Lee, 2006). Therefore, to find the practical way to improve western audiences' reception of Chinese films, this study will explore the following three questions:

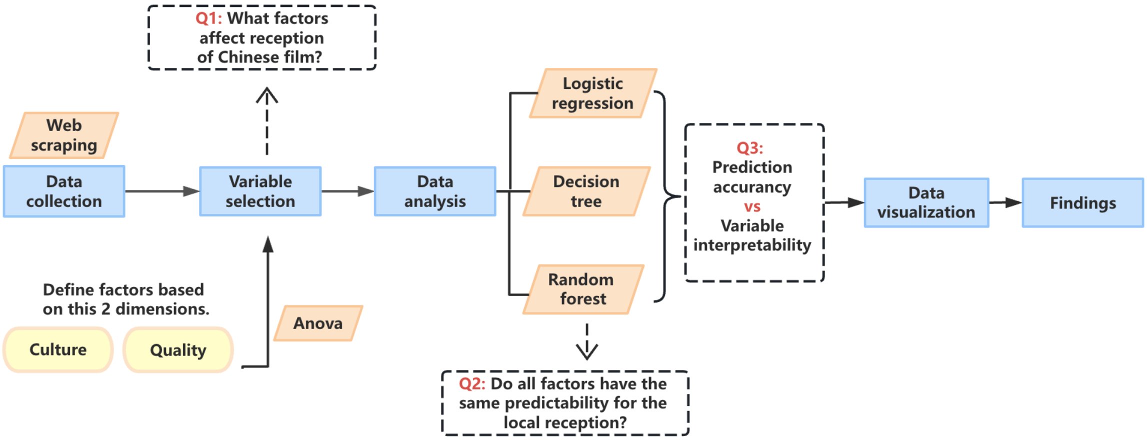
Q1: *What factors affect the reception of Chinese films in western markets?*
Q2: *Do all factors have the same predictability to the local reception? If not, which factors are more important? And how do they affect the reception?*
Q3: *Is the local reception of Chinese films in western market predictable with machine learning models? If predictable, what is the best model?*

1.3 Research organization

Figure 1 depicts the research structure of this study. The study primarily utilizes IMDb.com and Douban.com as data sources and applies web scraping techniques using R programming to collect data. The collected data undergoes data wrangling, wherein the factors are defined based on two dimensions, namely culture and quality. ANOVA is employed to determine the significant variables that affect the reception of Chinese films by western audiences. Logistic Regression models, Decision Tree models, and Random Forest models are subsequently developed with the significant variables and optimized based on prediction accuracy with a Cross-Validation method. Finally, the results of quantitative analysis and data visualization are combined.

Figure 1

Research structure


Q1: What factors affect reception of Chinese film?
Web scraping
Data collection
Variable selection
Data analysis
Logistic regression
Decision tree
Random forest
Q3: Prediction accurancy vs Variable interpretability
Data visualization
Findings
Define factors based on this 2 dimensions.
Culture
Quality
Anova
Q2: Do all factors have the same predictability for the local reception?

The chapter structure of this study is as follows: Chapter 1 provides the background and context of the study, introducing the research questions. Chapter 2 summarizes the current state of research, proposes hypotheses based on existing theory, and discusses their implications. Chapter 3 explains the data collection and research methods. Chapter 4 outlines the data processing steps, model building procedures, and presents the analysis results. In Chapter 5, the findings of this study are compared and discussed with existing research. Finally, Chapter 6 describes potential applications of the study's results and suggests future improvements for analysis in this topic.

CHAPTER 2. Literature Review

2.1 Relavant concept

Local reception has long been studied as a problem of meaning construction with the use of qualitative research methods. ***Local reception***, broadly understood as the processes in which meanings and values of media products are created within each local background, usually has been the subject of qualitative studies. It was first time to be applied in an empirical analysis with two quantitative manifestations: cultural discount and cross-culture performance predictability in Lee’s research (2006).

The term "***cultural discount***" was coined by Hoskins and Mirus (1988) to explain the dominance of the US in the global television programming trade. The theory of cultural discount, which is popular in media and cultural economics, suggests that a large cultural distance between the product-producing and product-consuming markets can reduce sales of the global product, because consumers in the product-consuming market may not fully appreciate the product's cultural content (Thompson & Chmura, 2015). In other words, cultural discounting can be viewed as a form of value friction that affects the consumption of cultural products, such as films, music, books, and more (Jane, 2021). Reducing culture distance is a key point to improve foreign audiences’ reception in the film industry.

Another concept, ***Cross-culture performance predictability***, refers to the ability to predict the performance of media products in one culture based on their performance in another culture, involving evaluating the consistency of performance across different cultural contexts for the same products (Lee, 2006). The original media product performance can not guarantee the same performance in a foreign market.

Although cultural discount and cross-culture predictability are closely related, it is beneficial to consider them as distinct concepts. Both are based on the recognition of cultural differences and may be mitigated by the universalization of media products. However, cultural discount pertains to the loss of values, whereas cross-culture predictability concerns the loss of predictability in values, and these two phenomena have distinct empirical manifestations (Lee, 2006).

2.2 Theory framework

In previous studies, the relationship between cultural factors and film performance has been widely evaluated in international markets. Craig et al. (2005) indicated that U.S. films perform better in countries that are culturally closer to the United States and with a greater degree of Americanization. The cultural similarity can strengthen the film performance in a cross-cultural market. It has also been proved by the study of Fu and Sim (2010) in a wider market, which revealed that film trade flow increase can be eroded by cultural distance between both countries, and the cultural discount can be moderated by domestic market size and sharing common language in both markets. Ye et al. (2018) indicated that the absence of systematic and persistent correlation in cultural distances and genre preferences from country to country. So the result of cultural difference analysis based on genre can not be completely applied in different markets. In recent study from Wang et al. (2021), the results concluded that cultural distance negatively influences on foreign box office revenue, and content specificity (the message of a film reflected in the story, e.g. "punchline" of comedy films) and aesthetics specificity (elements affect how the story is told) of films strengthens and weakens this relationship respectively. These studies all share a common view that cultural differences are an important factor on film performance in foreign markets, and universalization and high culture compatibility is the key solution to mitigate the culture discount.

The research mentioned strengthened the impact of culture discount on media product performance in cross-cultural markets. However, to date, research on the quantification and prediction of cultural discounts is very limited. Lee was first to quantify cultural discount by box office ratio of U.S. films in East Asian markets. He conducted three research through descriptive analysis and regression analysis to evaluate cultural discount and cross-cultural predictability to film performance difference

Lee's (2006) initial research focused on Hong Kong, analyzing the impact of genres on cultural discounts and predictability. The results showed that comedies were highly particularistic, and science fiction was the most universal. Then, Lee (2008) expanded the research to seven East Asian countries and found that comedy was quintessentially culturally specific, while adventures were universal. The results from both analyses are showing coexistence in the comedy genre. However, box office performances in cross-cultural markets are affected not only by cultural differences but also by movies' quality. Later, Lee (2009) added the "Award" variable to represent cinematic quality and artistic achievement and concluded that drama award films were relatively easier to be discounted by cultural differences. Overall, Lee's research primarily focused on the impact of cultural discounts and film quality on local reception, as represented by the genres and awards' effects on the box office ratio. Other studies have concluded that cultural differences and film quality are affected by various factors. It is clearly not sufficient to predict local reception of a film only including film genre and award as predictors.

2.3 Research gap

The previous research has examined in detail the causes of cultural discount, such as economic and cultural differences between trading markets; and how the local reception of films in cross-cultural markets is influenced by culture and film quality factors; and also how to predict local reception. However, the following three research gaps have not been covered by previous studies.

Lack of research on Chinese films in western markets: While a considerable body of research has been carried out on the western film reception in global markets (Craig et al., 2005; Lee, 2006, 2008, 2009; Moon et al., 2016; Gao et al., 2020), despite Wang et al. (2021) analyzed East Asian films performance in European markets, it is still less known about the others study on East Asian films, and even no Chinese film reception in western perspective, especially the films from China mainland.

Evaluate cultural discounts only based on box office: Previous studies mainly focused on comparing the box office performance to evaluate cultural discounts. There are two success-related aspects of a film: critical ratings and box office performance. The former indicates artistic excellence, and the latter reflects a movie’s commercial appeal (Holbrook & Addis, 2008). The same film box offices in different regions show differences in population size, ticket prices, length of the film release period, other general economic factors, and so on. Hence, the box office is not better than the review rating to evaluate cultural discounts.

Lack of Machine Learning based analysis on Cultural discount predictability: Most of the analysis on cultural discount and audience reception prediction is based on qualitative analysis. Even though, in recent years, quantitative analysis has performed on this topic, the research is mainly based on descriptive analysis and regression models(e.g. Lee, 2006, 2008, 2009; Wang et al., 2021).

To bridge the gap of the previous studies, this research will focus on consumer reception of Chinese films in western markets. Also, film ratings will be used to evaluate western people’s reception of a Chinese film, because film ratings can reflect films cultural and artistic values and are more comparable than film box office. Western audiences’ reception will be analyzed and predicted with Machine Learning models in this study.

2.4 Hypotheses development

To answer the research question of this paper: how to improve western audiences' reception of Chinese films, corresponding hypotheses are proposed for each sub-question based on the previous findings and will be tested later in this study.

Q1: What factors affect the reception of Chinese films in western markets?
H1: *Western audiences' reception of Chinese films may be affected by factors from two dimensions: cultural and film quality.*

High cultural similarity between two countries gains more film success. Craig et al. (2005) indicated that U.S. film box office receipts perform better in countries that are culturally closer to the United States and with a greater degree of Americanization. The cultural difference between cross-cultural markets is not the only factor that affects local reception. Since film quality is also important for domestic audiences' reception (Saraee et al., 2004). Lee (2009) analyzed the impact of both cultural factor and film quality factor on U.S. films local reception in East Asian markets, concluding the genre and award can be affected by cultural differences.

Regarding cultural similarity in the film industry, 4 factors are mostly discussed as the factors that affect cultural differences, that are film genre, language similarity, title informativeness, and star effect. People prefer movies and shows made in their native language. Although foreign language productions can be subtitled or dubbed, something is always lost in the translation process (Fu & Sim, 2010). The preferences of films with different genres are shown across countries(Craig et al., 2005; Lee, 2006), which means people from different countries have different genre preferences. Also, higher similarity with the original title and more informativeness of film title translation could provide the film more success in a foreign market (Gao et al., 2020). Saraee et al. (2004) strengths the star effect on film success.

Lee (2009) included the Academy Award to represent film quality in local reception prediction, resulting that drama awards relate negatively to local reception. Domestic success does not guarantee international success, or at least not to the same extent. The domestic success of a media product is a useful predictor of its international success only to a certain extent (Lee, 2006). Even though the previous study stands that domestic success is not equal to foreign market success, the study didn't deny the impact from domestic success on foreign market performance. Also, film production budget is often related to the film shooting and processing techniques, and is an important variable related to the quality of the film.

Thereby, in terms of both culture and film quality, the following factors may have an impact on local reception: genre, language similarity, title informativeness, star effect, award, domestic success and budget.

Q2: Do all factors have the same predictability for the local reception? If not, which factors are more important? And how do they affect the reception?
H2: The factors are not with the same predictability.

The predictability may be different across genre types. According to Lee's study (2006; 2008; 2009), comedies show good performance, while dramas do not. Similar language, title informativeness, star effect, award, domestic success and budget may bring positive impact on western audiences' reception. Although all seven factors mentioned in H1 have shown an impact on film performance in cross-cultural markets in previous studies, there may be some factors that have no impact on the western market for Chinese films, because the previous studies are performed on different markets. Therefore, the factors may show different predictability.

Q3: Is the local reception of Chinese films in western market predictable with machine learning models? If predictable, what is the best model?
H3: ***The local reception of Chinese films in western market is predictable. The best prediction model candidates can be Logistic Regression, Decision Tree and Random Forest models.***

In previous studies, researchers have compared some statistical and machine learning models' effectiveness on the popularity prediction of movies based on IMDb database. The analysis result from Latif and Afzal (2016) proved simple logistic and logistic regression achieved better prediction accuracy around 84% than Multilayer Perceptron, J48, Naive Bayes and PART models. Abidi et al. (2020) implemented the Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Tree models to predict the box office, and the result indicated Decision Tree's effectiveness on this topic. Later, in the study from Oyewola and Dada (2022), Bagging showed higher prediction accuracy than Multinomial Logistic Regression, Support Vector Machine, Naive Bayes and K-Nearest

Neighbor. Therefore, the Logistic Regression model, Decision Tree model and Random Forest model can be used to predict local reception.

To sum up, three machine learning models, Logistic Regression model, Decision Tree model and Random Forest model, can be conducted to explain what factors affect western audiences' reception of Chinese film. The factors from cultural dimension and film quality dimension are involved, including genre, language, title, star effect, budget, award, and Douban rating.

CHAPTER 3. Methodology

3.1 Data sources and data collection

3.1.1 Data selection

In this study, the film information is extracted from IMDb.com and douban.com databases, which are widely used in film-related academic research. The comprehensive attribute categories of IMDb.com make its advanced search function more professional and accurate. In addition, the star effect of actors is assessed by the top 1000 stars in leading roles at the worldwide box office from the the-numbers.com website; film awards are obtained from Wikipedia data (only the Oscars Academy Awards and the Golden Palm Awards are considered in this study).

To get the dataset that fits this research, the entire IMDb database was filtered in detail as follows: (1) Only feature films were analyzed. (2) Only films marked as China, including the Chinese mainland, Hong Kong, and Taiwan, in the database were analyzed. For some films in the search results that had Chinese companies involved but were not produced by China and whose content and actors were not related to China, this study judged that they did not meet the sample criteria, and therefore did a manual selection of each sample. (3) No restriction on language type. The reason is that the IMDb database has complex attributes under "Language", including Chinese, Mandarin, Cantonese, and Shanghainese. (4) The range of film release years is from 1970 to 2021.

3.1.2 Data collection

Using the filter rules outlined in section 3.1.1, relevant data were extracted by web scraping using the Rvest package in R. This involved downloading web pages from IMDb.com and Douban.com using the "read_html( )" function, extracting information using a loop across multiple pages, and storing the collected data using the "html_text( )" function. A sample of 2519 raw data points was collected on November 15, 2022, and summarized in Table 1.

Table 1

Raw data

Variable Name	Description	Missing Value	Data Resource
IMDb Rating	IMDb rating scores of a movie.	0	IMDb.com
Budget	Estimated budget of a movie.	2109	
Language	Available language of a movie.	0	
Runtime	Runtime of a movie.	23	
Genre	Genre of a movie	1	
Release Year	Release year of a movie.	0	
Original Title	The pinyin title of the film.	1	
Primary Title	The title is used in western markets, in a pinyin or English-translated version.	1	
Origin	Original region of a film.	0	
Douban Rating	Douban rating scores of a movie.	165	Douban.com
Rate Number	Number of Douban users rated the film.	165	
Award	Whether the movie or a cast member was nominated or won an Oscar Award or Golden Palm Award.	0	Wikipedia
Star Effect	The number of the top 1000 stars at the worldwide box office performed in a film.	0	the-numbers.com

3.2 Research methods

In this study, three machine learning models - Logistic Regression, Decision Tree, and Random Forest - were used for quantitative analysis. Data visualization was performed to enhance the validation of the results and extract more insights from the dataset. Sections 3.2.2

and 3.2.3 address the challenges of feature selection and overfitting in machine learning analysis, respectively. The methods used to solve these challenges are also presented.

3.2.1 Conceptual modeling

(1) Logistic Regression model

Logistic Regression model is used to analyze the relationship between a categorical dependent variable and one or more independent variables, and predict the probability of the occurrence of a binary event based on the values of the predictors. Logistic Regression models the relationship between the dependent variable and the independent variables using a logistic function, which produces an S-shaped curve that maps the values of the predictors onto the probability of the binary event. The output of Logistic Regression is a predicted probability that the binary event will occur, which can be used to classify the observations into one of two categories based on a threshold value.

The logistic function maps any real input *t* to an output value ranging between zero and one. It is commonly used in statistics to convert log-odds into probabilities. The standard logistic function, denoted as σ(*t*), is defined in function (1). If *t* is assumed as a linear function of a single explanatory variable *x*, *x* and general logistic function can be represented by function (2). So the Logistic Regression model with more explanatory variables can be represented by function (3).

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \tag{1}$$

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{2}$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{3}$$

(2) Decision Tree classification model

Decision Tree builds classification or regression models in the form of a tree structure, a top-down, greedy search through the space of possible branches with no backtracking. It partitions the data set into subsets that contain observations with similar values (homogenous). An associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A *decision node* has two or more branches. *Leaf nodes* are in the bottom layer representing a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called *root node*.

Deciding each node from different attribute orders will result in different tree structures. In order to make the final tree model simpler and more efficient, the attribute with the best classification ability on each decision node is selected. There are three different criteria for selecting the nodes' "best" attributes: Information gain, gain ratio, and Gini index. Both Information gain and gain ratio are used for classification trees. Gini index can be used as both classification trees and regression trees. For classification purposes, this study applies CART method using the Gini index as the basis. Gini index aims to decrease the impurities from the root nodes to the leaf nodes in a range of 0 to 1, where 0 represents the sample is pure on the nodes and vice versa.

In R programming, the function "rpart( )" is applied to CART method as default. Pruning is performed by adjusting the “cp” parameters, which is complexity parameter deciding minimum error improvement needed at each node to split. Use “plotcp( )” function to decide the optimal “cp” value, which applies 10-fold cross-validation by default.

(3) Random Forest classification model

Random Forest is an ensemble method built on the Decision Tree algorithm. It applies the *bootstrap aggregating* (or *Bagging*) technique to tree learners. *Bagging* repeatedly selects a random sample with replacement of the training set and fits trees to these samples. The trees are without pruning and independent of each other. Predictions can be obtained from each sample by majority vote for classification cases.

Random Forest can also be used to rank variable importance. There are two measures of importance provided for each variable. The first measure calculates the reduction in accuracy when a variable is removed. The second measure is based on the decrease in Gini impurity when a variable is chosen to split a node.

This study applies the Random Forest method based on CART decision tree. Use “randomForest” package in R. The "ntree" and "mtry" parameters are tuned using the cross-validation method to determine the optimal number of trees and the variables to sample as candidates at each split.

3.2.2 Feature selection methods

Guyon and Elisseeff (2003) summarized variable and feature selection methods including: dimensionality reduction (e.g. PCA), filter methods (correlation check, ANOVA, Chi-2 square, mutual information), wrapper methods (forward, backward selection), embedded methods (penalty regularization), and feature importances in tree base models. The methods have different evaluation criteria to select features. Some of the methods, like wrapper methods, are supervised methods that are evaluated based on the performance of a resulting model on a holdout dataset; some are statistical-based methods, like the filter methods, evaluating the relevance of each input variable outside of the models using statistics and selecting those variables that have the strongest relationship with the response variable (Kuhn & Johnson, 2013). Hence, to answer research question 1 mentioned in Chapter 1, perform correlation checks between each variable, remove the features that are irrelevant to the response variable and are with collinearity; then, apply ANOVA for feature selection.

ANOVA is a statistical test used to estimate the changes of the dependent variable in response to independent variables. Specifically, it examines whether there are differences in the means of the groups at each level of the independent variable. The null hypothesis (H0) of ANOVA is that there is no difference in means, while the alternative hypothesis (Ha) is that the means are different from one another. The criterion for identifying the “best” candidate features is that the P-value of the candidate has to be lower than 0.05 on an ANOVA F-test, which means the null hypothesis is rejected.

3.2.3 Overfitting problem

Overfitting is a key challenge in supervised machine learning, as it limits our ability to generalize models to fit both the observed data in the training set and the unseen data in the testing set. Overfitting can occur due to various factors, including the presence of noise, the small size of the training set, and the complexity of the model.

To mitigate the effects of overfitting, several strategies have been proposed to address its underlying causes. These strategies include: "Early-stopping" strategy, which prevents overfitting by halting training before the model's performance ceases to improve; "Network-reduction" strategy, which eliminates noise in the training set; "Data-expansion" strategy, which fine-tunes hyperparameters for complex models using a large amount of data; "Regularization" strategy, which ensures robust model performance by selecting useful features and distinguishing between more and less important features (Ying, 2019).

Cross-Validation is a commonly used technique to avoid overfitting in machine learning models. Cross-validation involves splitting the data into multiple subsets, or "folds", and using one fold as the test set while the rest of the folds are used as the training set. This process is repeated multiple times, with different folds used as the test set each time. By evaluating the model performance on multiple test sets, cross-validation helps to assess how well the model generalizes to new data and can help identify if the model is overfitting to the training data.

To avoid overfitting, this study will remove irrelevant features, lower models' complexity, and training models with Cross-Validation methods. All the Cross-Validation involved in this study are processed with 10 folds and 3 repeat times.

3.2.4 Summary

Research sub-question 1 will be answered by performing feature selection methods, which will give out the factors with significant effect on local receptions. Then, apply all selected variables into the Logistic Regression model, Decision Tree model and Random Forest model. The coefficient estimation and significance in Logistic Regression model, the selected variables in Decision Tree, and the variable importance in Random Forest model can indicate the factors' predictability difference on local reception, so the sub-question 2 will be answered. Research sub-question 3 can be solved by comparing the different models prediction performance and the variable interpretability.

When the three sub-questions are solved in turn, we can know which variables have a significant effect on western local reception, so that we can suggest improvements to the filmmakers to gain the acceptance of Chinese films in the western market; in addition, after obtaining the prediction model, we can also determine whether the film will be accepted by the western audiences and thus decide whether the film will be released in the western market.

CHAPTER 4. Data Analysis and Results

4.1 Data wrangling

4.1.1 Missing value

Missing values existing in raw data (Table 1) are processed in the following ways:

- For “Budget”, since 2109 out of 2519 data points in the dataset are missing the "budget" variable, remove this variable.
- Use the average runtime to fill in the missing values in the “Runtime” variable.
- There are 165 missing values in the “Douban Rating”. As the key feature to calculate the reception of a film, remove all the missing value samples in this variable.
- As there is only 1 missing value in each of “Genre”, “Original title”, and “Primary Title”. Remove the samples with missing value.

4.1.2 Variable definition

(1) Dependent variable

The western audiences’ reception of Chinese films can be defined as the difference between the rating score on IMDB and Douban. Delete all the samples without “Douban Rating " in the raw dataset. IMDb and Douban define the rating mechanism as 1-10 stars and 1-5 stars respectively. One star means the least satisfaction, and ten and five stars indicate the most satisfaction on both platforms respectively.

To remove the rating system difference, this research defines 10 as a full rating score for both platforms, which means the IMDb rating mechanism is 1-10 stars and the Douban is with a 2-10 stars rating system (i.e., 2, 4, 6, 8, 10). Then, set 0.67 points as a threshold to define the western audiences' reception of a Chinese film. To specify this, when the difference between the ratings of the same film on the two platforms is within 1 point, it can be considered the same rating score on both platforms. For example, if the Douban rating is 6, which could be the same meaning of IMDb rating as 5, 6 or 7, because of the lack of 5 and 7 score in Douban rating scale. The probability of Douban rating is shown with the same score on IMDb is 33.3% , P(same) = ⅓, and shown as 1 score difference is 66.7% , P(diff) = ⅔. The threshold to evaluate the rating difference between Douban and IMDb is 0.67. The calculation algorithm is shown in the function below.

$$Threshold = 1 * P(diff) + 0 * P(same)$$

Therefore, a western audience accepting a Chinese movie is represented by the same or higher rating score on IMDb, which means the rating difference should be equal to or higher than 0.67. Create a new variable named "Accept", with a level of "Yes" to represent reception, and a level of "No" to represent non-reception.

(2) Dependent variables

"English": Language diversity and different language preferences are showing in western markets (e.g. America and Europe). Since English is the most well-used language in western markets, only English is taken into account to evaluate the language effect on western audiences' reception of Chinese films. English version availability of a Chinese film is defined as variable "English" with two levels "English" and "No English", indicating English is included or not in "Language" from the raw dataset.

"Genre": Films are not only limited to a single genre on IMDb.com. A film can belong to one or many genre types. For example, *Shaolin and Wu Tang* released in 1983 has a single genre of action, whereas the film *The Way of the Dragon* released in 1972 has 3 genres that are action, adventure, and comedy. The variable "Genre" originally includes 15 levels, which are independent of each other. To represent the genre for each film, 15 dummy variables are assigned to each film: “Action”, “Adventure”, “Biography”, “Comedy”, “Crime”, “Drama”, “Family”, “Fantasy”, “History”, “Horror”, “Music”, “Romance”, “Sport”, “War” and “Western”.

“Title”: The titles of Chinese films rendered in Pinyin, which is a phonetic writing system for Modern Standard Chinese, lower the efficiency of western audiences comprehending key information about the films. This problem can be mitigated if the primary title of the film is translated into English. To deep dive into the effect of title version difference on audiences’ reception, a new variable called "Title" is created, which is based on a comparison between the "Original Title" and the "Primary Title" of a film in the raw dataset. If the "Original Title" and the "Primary Title" are the same, then the "Title" variable will be defined as "Pinyin". Conversely, if the "Original Title" and the "Primary Title" are different, then the "Title" variable will be defined as "English".

4.1.3 Final dataset overview

This research includes a balanced dataset of 2354 samples, with 1121 labeled as "Accept = Yes" and 1233 labeled as "Accept = No". The dataset features 27 variables without missing values. The final dataset was randomly split into training and testing sets at a 7:3 ratio, setting seed (16666) for reproducibility. A detailed data description is provided in Table 2.

Table 2

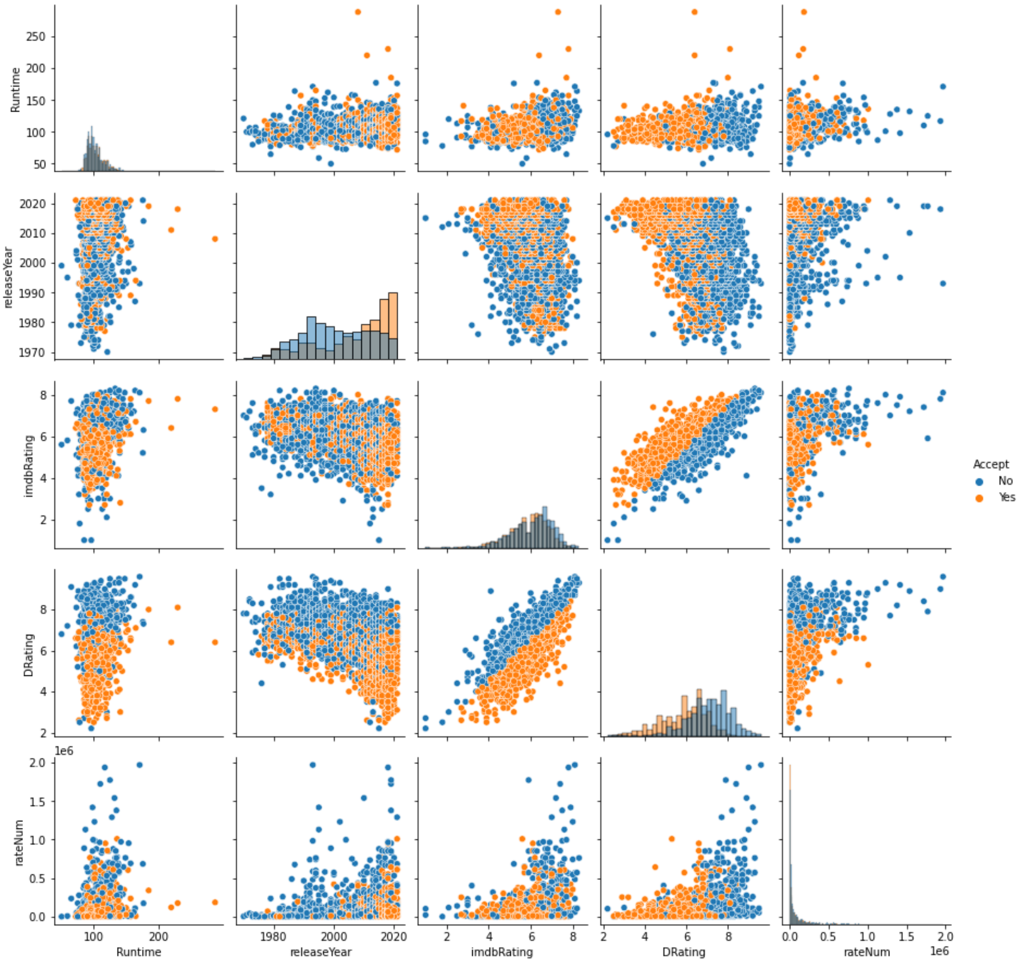
Final dataset

Variable	Description	Data Type
Accept	"Yes": The film is accepted by western audiences. "No": The film is not accepted.	Character
imdbRating	IMDb rating scores of a film.	Numeric
English	"English": English version is available for a film. "No English": English version is not available.	Character
Runtime	Runtime of a film. (minutes)	Numeric
Genre: 16 dummy variables	"1": a film belongs to the genre type. "0": a film doesn't belong to the genre type. Film genre variables include: "Action", "Adventure", "Animation", "Biography", "Comedy", "Crime", "Drama", "Family", "Fantasy", "History", "Horror", "Music", "Romance", "Sport', "War", "western".	Character
releaseYear	Release year of a movie.	Integer
DRating	Douban rating scores of a film.	Numeric
rateNum	Number of Douban users rated the film.	Integer
Title	"English": The title of the film is translated into English "Pinyin": The title of the film is in Chinese pinyin.	Character
Origin	Original region of a film: "Mainland", "Hongkong", and "Taiwan".	Character
Award	"Award": The film was nominated or won an Oscar Award or Golden Palm Award. "No Award": No western Award.	Character
starEffect	The number of stars performed in the film.	Integer

4.2 Feature selection

Figure 2

Par plot of variables



A parallel coordinate plot (par plot) was utilized to visualize the interrelationships between the variables. As shown in the par plot, a collinearity issue is observed between the variables "DRating" and "imdbRating". The correlation coefficient calculated between these two variables is 0.81, indicating a high positive correlation between them. This observation is consistent with Lee's previous research (2006), which demonstrated that a media product's success in its original market is a strong predictor of its international success. Consequently, "DRating" was retained while "imdbRating" was removed from the original dataset.

A logistic regression model was developed using the remaining variables in the dataset, and the ANOVA method was employed for variable selection. The significant features identified through ANOVA are presented in Table 3.

4.3 Model result

Apply all selected variables to the logistic regression model, decision tree model, and random forest model. Then, train the models and select the model with the smallest Cross-Validated prediction error as the optimal model. Finally, the testing data set is applied to the final models, comparing the accuracy of the three prediction results to determine the optimal model method.

(1) Logistic regression model

As per the ANOVA analysis, logistic regression models were developed using variables including "Runtime", "releaseYear", "DRating", "rateNum", "Action", "Animation", "Comedy", "Drama", and "Music". The models were then subjected to 10-fold cross-validation, repeated thrice on the training dataset, to determine the average prediction accuracy. A logistic model was then executed on the entire training set, with the coefficient estimation provided in Model 1. Variables with no significant impact, as indicated in Model 1, were eliminated, and the remaining variables were subjected to another round of cross-validation to estimate the prediction accuracy. Model 2 was then developed using the complete training dataset to determine the variable coefficients. This process was repeated, and the results of cross-validated prediction accuracy and coefficient estimation for Model 3 and Model 4 are presented in Table 3.

Upon comparing the prediction accuracy of the various models, it was determined that Model 3 exhibited the highest accuracy during cross-validation. This model was selected as the optimal logistic model and subsequently tested with the testing dataset, achieving an accuracy of 75.81%, sensitivity of 79.68%, and specificity of 71.47%. The AUC of the prediction is 0.837, which means the prediction performs well on both sensitivity and specificity.

The optimal logistic regression model indicates that "DRating", "Runtime", "Comedy", "Drama", "Action", and "Music" significantly impact the reception of Chinese films among western audiences with a 95% confidence level. A one-point increase in the "DRating" variable corresponds to a 74.35% decrease in the odds of accepting a film. Moreover, a one-minute increase in "Runtime" corresponds to a 1.51% increase in the odds of accepting a film. Among the genre variables, action films have 33.51% higher odds of being accepted compared to non-action films, while comedy films have 39.77% lower odds of being accepted compared to non-comedy films. In contrast, drama films have 80.58% higher odds of being accepted compared to non-drama films, and music films have 188.93% higher odds of being accepted compared to non-music films. Based on these results, it can be concluded that variables such as "DRating", "Drama", and "Music" exhibit higher predictability regarding the reception of Chinese films by consumers, compared to other variables.

Table 3

Anova and logistic regression model results

Variables	Anova (P-value)	Logistic model coefficient estimation (P-value) and model performance			
		Model 1	Model 2	Model 3	Model 4
(intercept)	NULL	-24.160 . (0.0774)	-13.025 (0.2788)	6.988 *** (< 2e-16)	7.018 *** (<2e-16)
Runtime	(0.0304) *	0.012 * (0.0104)	0.012 ** (0.0093)	0.015 *** (0.0004)	0.015 *** (0.0003)
releaseYear	(< 2.2e-16) ***	0.015 * (0.0229)	0.010 . (0.0959)		
DRating	(< 2.2e-16) ***	-1.259 *** (<2e-16)	-1.308 *** (< 2e-16)	-1.356 *** (< 2e-16)	-1.337 *** (<2e-16)
rateNum	(0.0286) *	-6.337e-07 (0.2350)			
Action1	(0.0063) **	0.389 ** (0.0045)	0.356 ** (0.0083)	0.289 * (0.0241)	
Animation1	(0.0031) **	-0.690 (0.1107)			
Comedy1	(5.247e-06) ***	-0.469 *** (0.0008)	-0.485 *** (0.0004)	-0.507 *** (0.0002)	-0.544 *** (5.26e-05)
Drama1	(0.0002) ***	0.550 *** (0.0001)	0.581 *** (3.94e-05)	0.591 *** (2.85e-05)	0.508 *** (0.0002)
Music 1	((0.0415) *	1.006 * (0.0396)	1.031 * (0.0353)	1.061 * (0.0312)	
Mean prediction accuracy on training data set with cross-validation		75.25%	75.39%	75.63%	75.11%

*Note: Signif. codes 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

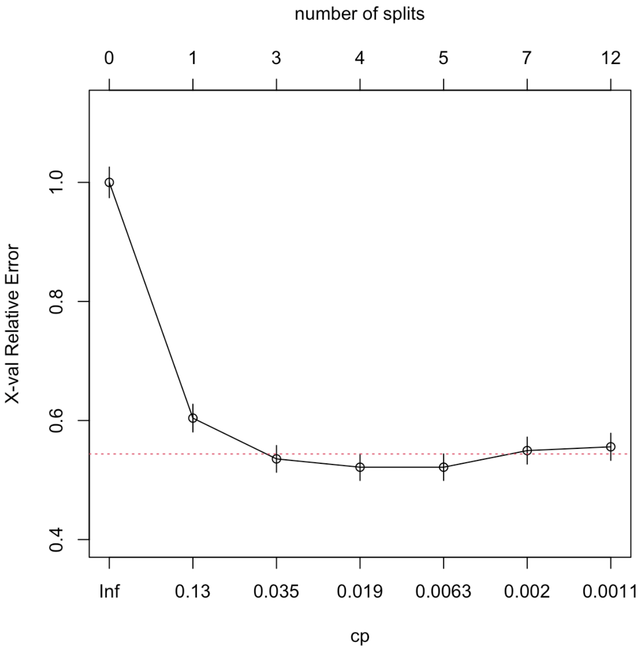
(2) Decision Tree

When building the decision tree model, it is easy to have overfitting problems. The simpler the structure of the tree is, the more robust the model will be. So the optimal decision tree model should satisfy the simplest structure in the case of having the optimal performance.

In R programming, the “rpart” implementation first fits a fully grown tree on the entire training data set. After this step, the tree is pruned through adjusting complexity parameter (CP) value. Use “plotcp” function to select the optimal complexity parameter value, which uses a 10-fold cross-validation method as default to plot cp values against the geometric mean to depict the standard deviation until the minimum value is reached. For every split, the validation error is expected to be reduced, but if the model is suffering from overfitting, the cross validation error increases. As the cross-validation result shown in Figure 3, the validation errors of CP equals 0.019 and 0.0063 are similar and both blow geometric mean. Therefore, select 0.019 as optimal CP value with less slits.

Figure 3

Complexity parameter value selection result

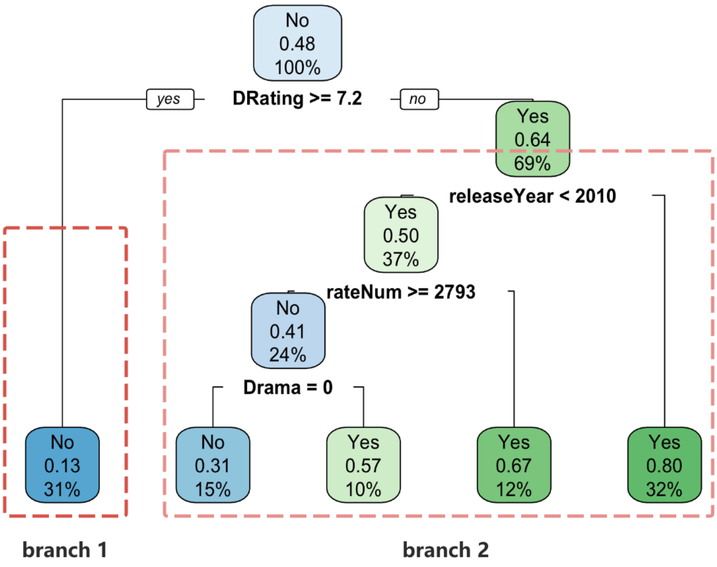


The best Decision Tree model is shown in Figure 4. The nodes from top to bottom present the importance of accordingly variables from high to low. The most important variables are “DRating”, “releaseYear”, and “rateNum” and “Drama”. In branch 1, the films with “DRating” higher than 7.2 are classified as not accepted by western audiences, the leaf node contains around 31% samples; branch 2 shows when “DRating” are lower than 7.2, only non-drama films whose "releaseYear" are not later than 2010 and "rateNum" are not less than

2793 are classified as not accepted containing around 15% samples, all the others are accepted by western audiences containing around 54% samples. Run the final decision tree model with testing data set, getting prediction accuracy 76.94%, sensitivity 76.20%, and specificity 77.78%. The AUC value is 0.806, showing this model with good prediction ability on both positive and negative results of the "Accept" variable.

Figure 4

Decision tree result



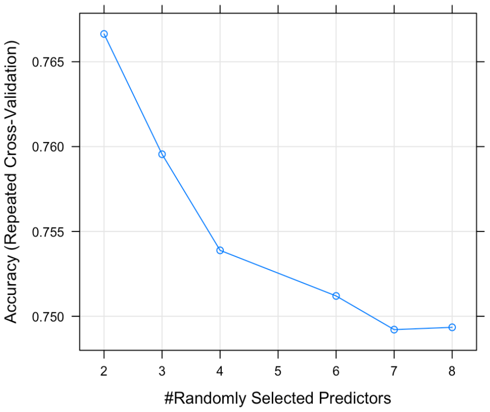
(3) Random Forest

To establish a baseline for comparison, a Random Forest model was created using all the variables selected in the ANOVA analysis. The default parameters "mtry = 7" and "ntree = 500" were used with the entire training set, resulting in an average cross-validated prediction accuracy of 75.2%. To refine the model parameters, which have a significant impact on the accuracy of the model, a 10-fold cross-validation method was applied.

Initially, various values of "mtry" were randomly tested within a range of 2 to 8. The cross-validated prediction accuracy performance is illustrated in Figure 5. The results revealed that "mtry = 2" yielded the most accurate prediction, with an accuracy of 76.6%. This value of "mtry" was then set, and the accuracy was compared for different values of "ntree" (i.e., 100, 200, 300, 400, and 500). The optimal number of trees was found to be 300, resulting in an accuracy of 76.6%.

Figure 5

Tune "mtry" result

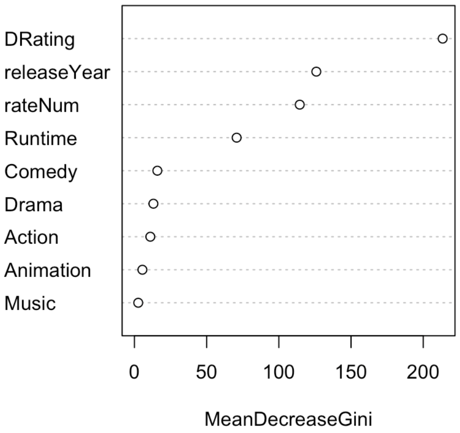


The variable importance plot (Figure 6), obtained from the optimal Random Forest model (mtry = 2, ntree = 300), is presented below. The variable importance algorithm used in the Random Forest model is based on the diversity of values that variables possess, which plays a crucial role in determining the variable importance. Only the numeric variables exhibit a high contribution towards decreasing the mean Gini index, thereby indicating their relative importance. The variable importance plot depicts that "DRating", "releaseYear", and "rateNum" are the most important numeric variables.

The final model was evaluated using the testing set, which resulted in a prediction accuracy of 77.37%, sensitivity of 76.47%, and specificity of 78.38%. Moreover, the AUC of the prediction result was calculated to be 0.844, signifying good predictive performance for both positive and negative results.

Figure 6

Variable importance



(4) Model comparison

The results of the testing set prediction using the optimal Logistic, Decision Tree, and Random Forest models are summarized in Table 4. All models achieved a prediction accuracy above 75%, with an AUC value greater than 0.8, suggesting that they all demonstrate good predictive ability. However, the Random Forest model exhibited the highest prediction performance in terms of the western audience reception of Chinese films.

Table 4

Summary of model results

	Sensitivity	Specificity	Accuracy	AUC
Logistic Regression	79.68%	71.47%	75.81%	0.837
Decision Tree	76.20%	77.78%	76.94%	0.806
Random Forest	76.47%	78.38%	77.37%	0.844

In terms of variable importance for predicting western audience reception of Chinese films, the three models used in this study showed different results. The Logistic Regression model identified "DRating", "Runtime", "Comedy", "Drama", "Action", and "Music" as significant predictors of “Accept”, with "DRating", "Drama", and "Music" exhibiting higher predictability; the Decision Tree model ranked “DRating”, “releaseYear”, “rateNum”, and “Drama” as the most important variables in predicting western consumer reception; while the Random Forest model identified “DRating”, “releaseYear”, and “rateNum” as the most important predictors.

Overall, the Random Forest model showed better prediction performance compared to the Logistic Regression and Decision Tree models. The variables “releaseYear”, “Runtime”, “rateNum”, “Comedy”, “Action”, and “Music” were found to have significant impact on the “Accept” variable in at least one of the models, while “DRating” and “Drama” were consistently identified as important predictors across all models.

4.4 Data visualization

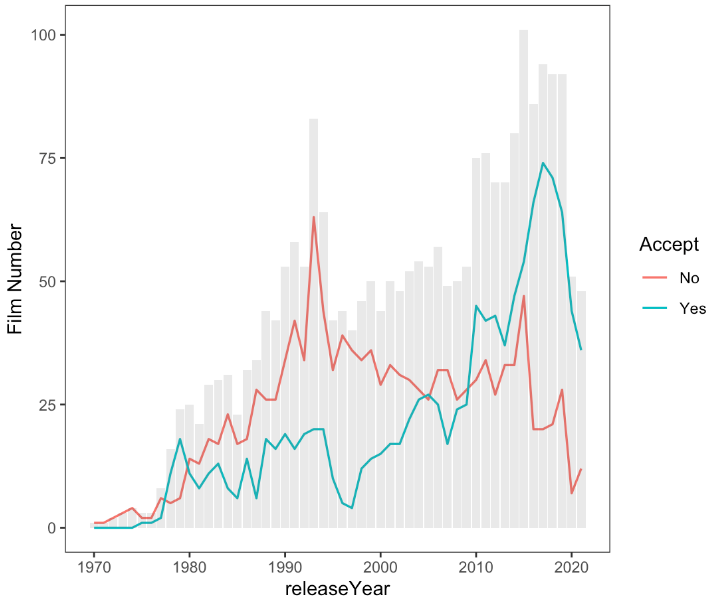
In this section, perform data visualization based on the model result to deep dive the relationship between variables and explore the further marketing insights related with cultural discount and film quality on West audience reception of Chinese films.

(1) Release year & Rating scores

In the figure below, bar plots represent the total number of Chinese films released each year, and line plots represent the number of films that were accepted by western audiences and the number that were not accepted each year. The number of Chinese film releases peaked in the 1990s and 2010s. In the 1990s, far fewer Chinese films were accepted by western audiences than those that were not, while the 2010s show the opposite result. This opposite situation may be due to technological differences between generations, which facilitate the western audience reception of Chinese films.

Figure 7

Number of annually released films and accepted films



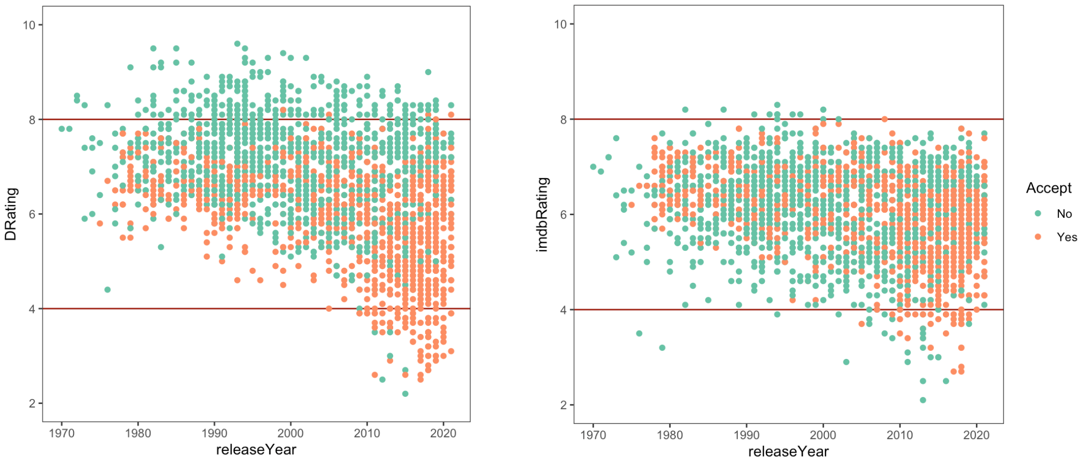
The Logistic model result indicates that a one-point increase in the Douban rating corresponds to a 74.35% decrease in the odds of accepting a film. The Decision Tree result shows 31% films that are with Douban rating higher than 7.2 are classified as not accepted, while the accepted films are mostly with Douban rating lower than 7.2, which is more significant on the movie released after 2010. High douban rating means the film is popular in

China with high quality. Models' result indicates the lower quality films are easier to be accepted. To deep dive into the reason behind this, draw the figures below showing rating distributions across release years on Douban and IMDb.

Comparing the rating distribution from Douban and IMDb, find that western audiences are more conservative than Chinese audiences in their ratings of Chinese films, with ratings mainly in the 4 to 8 range; Chinese audiences have more extreme ratings, especially for films with ratings below 4, which are found in a large number of films released after 2010. In the 1990s, films with high Douban ratings were slightly more than the other periods. Combining the findings from Figure 7, the conclusion is that cultural discount phenomena are more likely to exist on films with high Douban rating.

Figure 8

Distribution plots of Douban rating and IMDb rating



In the data set of this study, there are 17 films that were nominated or won an Oscar Award or Golden Palm Award. But only 4 films are accepted by western audiences. Therefore, the conclusion is that cultural discount phenomena are more likely to exist on films with high Douban rating, which is consistent with Logistic Regression and Decision Tree results.

(2) Drama

The Figure 9 below is about the rating scores distribution of Drama films and non-Drama films on Douban and IMDb platforms. The blue dots represent Drama films. Most of the Drama films on both platforms are showing with relatively higher rating scores. In Figure 10, it shows that Drama films with higher rating scores, especially on Douban, are more likely to get cultural discounts in cross-cultural markets.

Figure 9

Rating distribution of Drama films and Non-drama films

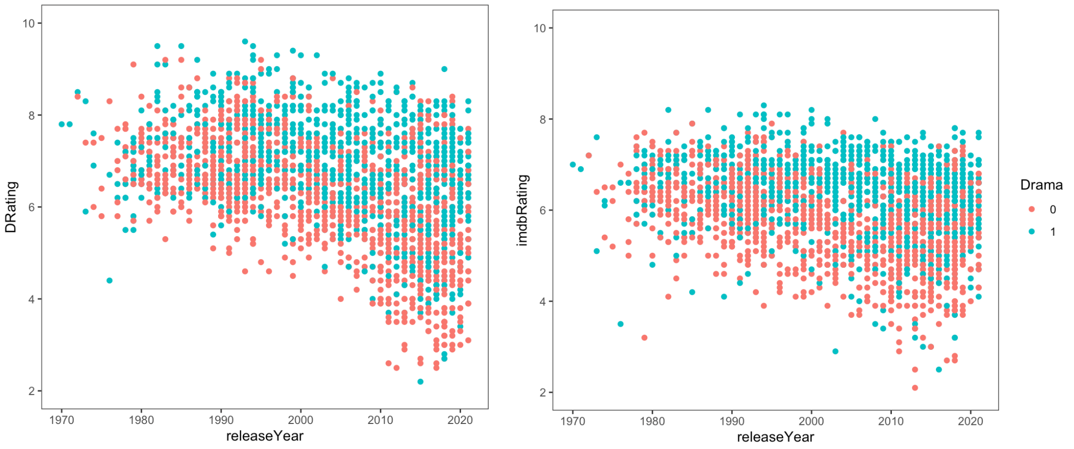
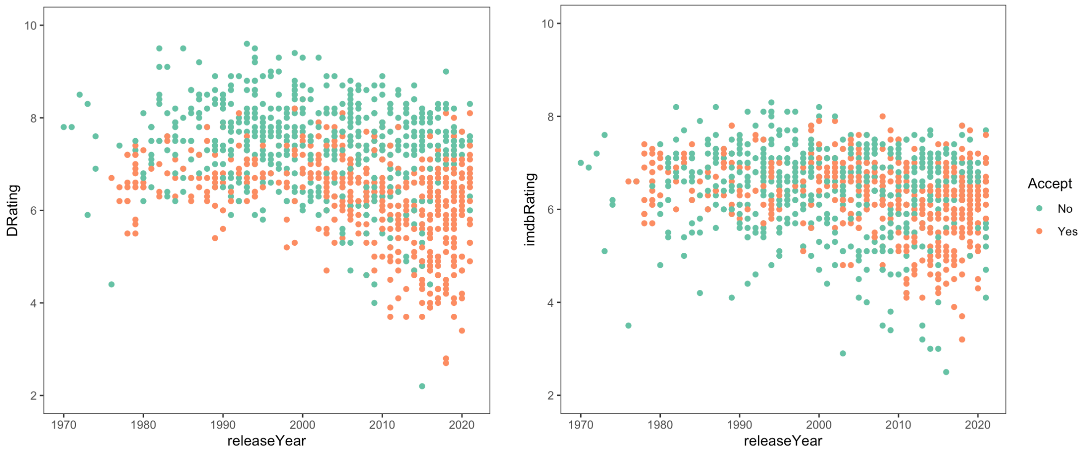


Figure 10

Drama film rating distribution plots



4.5 Findings

Based on the analysis presented, it can be concluded that important factors that affect the reception of Chinese films in western markets include Douban rating, release year, runtime, Douban rate number, and genre, particularly drama, comedy, action, and music. "DRating" and "Drama" have better predictability and also indicate the film quality and cultural factors have an impact on western audience reception.

There are two key findings about “DRating” and “Drama”. First, Chinese films with high Douban rating scores are always not accepted in western markets, facing cultural discount. While films with lower Douban rating are well accepted by western audiences, especially the film released after 2010. Second, drama films are with higher probability to be accepted in western markets than non-Drama films.

The Machine Learning models employed in this study, namely the Logistic Regression model, Decision Tree model, and Random Forest model, showed good prediction accuracy, all above 75%, with AUC values higher than 0.8. Although the Random Forest model had the highest prediction accuracy (77.37%), it prefers continuous variables for important variable selection and lacks interpretability. The Decision Tree model (76.94%) had a slightly lower prediction accuracy than Random Forest but much better than the Logistic Regression model (75.81%). Therefore, in terms of interpretability and prediction ability, the Decision Tree model is considered the best choice for this study.

CHAPTER 5. General Discussion

High Douban Rating films show lower local reception and more cultural discount in western markets. This conclusion is consistent with Elberse and Eliasberg study results (2003). They indicated culture differences may bring different reception levels, which means, for individual movies, domestic success does not always guarantee international success. There are two possible reasons behind this phenomenon: first, high douban rating films with more cultural discount may be caused by high quality movies involving more cultural specificity content; second, the different review rating habits between domestic audiences and western audiences, which has been shown in Figure 8. Therefore, the existence of cultural discount phenomena in a movie is not a basis for refusing to introduce the movie to foreign markets. For some high-scoring movies, even though there is a cultural discount, these movies can still perform well in foreign markets, so these high-quality movies are still worth being introduced to the western market.

Release Year of Chinese films has a significant impact on the acceptance of films in the western market. According to data visualization results, although the number of Chinese films released peaked in the 1990s and 2010s respectively, the acceptance of Chinese films in western markets has shown a contrasting picture. In previous quantitative studies on cultural discount, researchers mainly emphasized on the finance performances and have overlooked the impact of technology development and progress on film market performance over time. Since the 1990s and into the 2010s, the rise of Internet technology has facilitated cross-cultural communication among movie audiences, making them more receptive to foreign products.

As mentioned in the Motion Picture Association's (MPA) *2021 Theme Report* [4], the theoretical home/mobile entertainment market (content distributed digitally and on disc) has grown rapidly in the last three years, with a 14 percent increase compared to 2020, driven by digital. In 2021, the number of subscriptions to online video services worldwide rose by 14 percent from the previous year. As a result of a substantial increase of 26 percent, online video subscriptions have now become the second largest revenue market, overtaking satellite TV. With the widespread use of online video subscription platforms like Netflix, audiences will have greater access to foreign cultures, and cultural discounts can be mitigated by progressive cultural globalization; for movie investors, the rise of online video subscription has broadened the market for movies, making it possible to sell movies outside of cinemas; and also, for the movie industry to gain more consumer touchpoints.

The findings that drama films are more likely to be accepted in western markets, which is consistent with Lee’s (2009) analysis result. However, based on machine learning model results, genre factors are not the most important. It reveals that, compared with 10 years ago, more cultural universalization shows in the film industry. This can also be attributed to the technological advances that mitigate cultural differences.

Although some previous studies (Latif & Afzal, 2016; Abidi et al., 2020; Oyewola & Dada, 2022) have evaluated the model's prediction performance to assess the superiority of the model, this study considers not only the prediction ability of the model for local reception, but also the explanation ability of the model. So, in this study, Decision Tree outperformed Logistic Regression model and Random Forest model. In addition, considering that the prediction of movie reception will be widely used in the decision making of investors and movie makers in the foreign market, model replicability and robustness is also an important factor in evaluating the model. Therefore, this paper uses the cross-validation method to

[4] MPA 2021 Theme Report

pruning and tuning models, and sacrifices a small amount of prediction accuracy to select the simplest model as the optimal model to avoid overfitting that reduces the model's ability to explain new data.

CHAPTER 6. Conclusions

6.1 Implication

Some cultural differences that lead to cultural discount are not necessary to be removed by changing the movie itself, for example, Douban high-rated movies are more likely to encounter cultural discount. The Chinese film market can focus on the adaptations brought by technological development to the film industry, such as developing overseas online video subscription platforms, to make films with Chinese cultural characteristics more accessible to western audiences and reduce cultural barriers in the long run.

The research in this paper can also help investors to judge the performance of movies in foreign markets. For low-scoring movies in the Chinese market, the prediction model in this paper can be used to determine whether cultural discount appears or not. If there is no cultural discount, it means that the movie is suitable to be introduced into foreign markets and can bring more economic benefits in western markets Also, according to the findings mentioned above, drama film is a good opportunity for investors to expand the western markets for Chinese films.

6.2 Limitation and suggestion

In future studies, how to apply the review rating score to assess the effect of cultural differences on cultural reception still needs to be optimized. According to the definition of cultural reception, the presence of culture discounts indicates that a film is not accepted and vice versa. The key to quantifying local reception lies in the comparison of the difference in review rating between different markets. The Douban movie ratings applied in this paper are derived from the arithmetic average of all user ratings, but the IMDb rating mechanism is not yet publicly available. Therefore, while comparing the differences in ratings, this paper unifies the differences brought by the rating score scale and does not delve into the mechanism of the rating algorithm. The inaccuracy of local reception calculation may also cause the error in analysis. In future research, quantifying local reception with rating differences still needs to be discussed.

This study only applies to IMDb.com and Douban.com as the main data sources. Due to the absence of some data on these two platforms, such as budget and masterpiece adaptations, the influence of these variables on cultural differences cannot be judged in this study. In addition, according to the findings of this paper, technological advances have made Chinese movies more widely accepted by western audiences now than in the 1990s. To make the future research on this topic more comprehensive, the influence of online video subscription platforms can be included as a factor affecting the reception of Chinese films by western audiences.

Due to the limited analysis on quantifying cultural discount and local reception so far, this paper only studies three classical machine learning models, namely Logistic Regression model, Decision Tree model, and Random Forest model, in this topic. Therefore, it is recommended to try more machine learning models in the future research.

References

Abidi, S. M. R., Xu, Y., Ni, J., Wang, X., & Zhang, W. (2020). Popularity prediction of movies: from statistical modeling to machine learning techniques. Multimedia Tools and Applications, 79, 35583-35617.

Ang, I. (1989). Watching Dallas: Soap opera and the melodramatic imagination. Psychology Press.

Cantor, M., & J. Cantor (1986). American television in the international marketplace. Communication Research, 13(3), 509–520.

Craig, C. S., Greene, W. H., & Douglas, S. P. (2005). Culture matters: Consumer acceptance of US films in foreign markets. Journal of International Marketing, 13(4), 80-103.

De Vany, A. (2003). Hollywood economics: How extreme uncertainty shapes the film industry. Routledge.

Elberse, A., & Eliashberg, J. (2003). Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. Marketing science, 22(3), 329-354.

Fu, W. W., & Sim, C. (2010). Examining international country-to-country flow of theatrical films. Journal of Communication, 60(1), 120-143.

Gao, W., Ji, L., Liu, Y., & Sun, Q. (2020). Branding cultural products in international markets: a study of Hollywood movies in China. Journal of Marketing, 84(3), 86-105.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Holbrook, M. B., & Addis, M. (2008). Art versus commerce in the movie industry: A two-path model of motion-picture success. Journal of Cultural Economics, 32, 87-107.

Hoskins, C., & Mirus, R. (1988). Reasons for the US dominance of the international trade in television programmes. Media, Culture & Society, 10(4), 499-515.

Jane, W. J. (2021). Cultural distance in international films: An empirical investigation of a sample selection model. Journal of Economics and Business, 113, 105945.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

Latif, M. H., & Afzal, H. (2016). Prediction of movies popularity using machine learning techniques. International Journal of Computer Science and Network Security (IJCSNS), 16(8), 127.

Lee, F. L. (2006). Cultural discount and cross-culture predictability: Examining the box office performance of American movies in Hong Kong. Journal of Media Economics, 19(4), 259-278.

Lee, F. L. (2008). Hollywood movies in East Asia: Examining cultural discount and performance predictability at the box office. Asian journal of communication, 18(2), 117-136.

Lee, F. L. (2009). Cultural discount of cinematic achievement: The academy awards and US movies' East Asian box office. Journal of Cultural Economics, 33(4), 239-263.

Liebes, T., & Katz, E. (1986). Dallas and Genesis: Primordiality and seriality in popular culture.

Moon, S., Mishra, A., Mishra, H., & Kang, M. Y. (2016). Cultural and economic impacts on global cultural products: Evidence from US movies. Journal of International Marketing, 24(3), 78-97.

Oyewola, D. O., & Dada, E. G. (2022). Machine Learning Methods for Predicting the Popularity of Movies. Journal of Artificial Intelligence and Systems, 65-82.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

Rosen, S. (2021). Obstacles to using Chinese film to promote China's soft power: Some evidence from the North American market. Journal of Chinese Film Studies, 1(1), 205-221.

Saraee, M. H., White, S., & Eccleston, J. (2004). A data mining approach to analysis and prediction of movie ratings. Transactions of the Wessex Institute, 343-352.

Thompson, F. M., & Chmura, T. (2015). Loyalty programs in emerging and developed markets: the impact of cultural values on loyalty program choice. Journal of International Marketing, 23(3), 87-103.

Wang, X., Pan, H. R., Zhu, N., & Cai, S. (2021). East Asian films in the European market: The roles of cultural distance and cultural specificity. International Marketing Review, 38(4), 717-735.

Waterman, D. (2009). Hollywood's road to riches. Harvard University Press.

Ye, H., Binwei, L., & Starkey, G. (2018). Economic and cultural implications of china's one belt one road initiative for the film industry: cultural distance and taste preference. Australian Economic Papers, 57(3), 250-264.

Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.