The European Film Industry as a Market for Lemons

Antonios Lagarias
APP Lab
Rennes 2 University
35000 Rennes, France
antonios.lagarias@univ-rennes2.fr

Moritz Pfeifer
Institute of Economic Policy
University of Leipzig
04109 Leipzig, Germany
pfeifer@wifa.uni-leipzig.de

October 20, 2024

Abstract

This paper looks at the impact of film criticism on predicting the commercial success of European films. We focus on the EU-funded film journal *Cineuropa* and compare it against three large industry press outlets. Using sentiment analysis to evaluate film reviews and a binary classification algorithm to assess commercial success, we find that *Cineuropa* has a positive bias. This bias leads the prediction model to overestimate the likelihood of commercial success of films regardless of their actual market performance. We interpret this as evidence that *Cineuropa* functions more as an advertisement platform than as a source of film criticism, creating misleading signals independent of a film's artistic or commercial merit. This paper makes three contributions: (1) it presents the first predictive study on the European film industry as a whole, (2) it introduces a novel method for classifying commercial success with an emphasis on model explainability, and (3) it adds to empirical research on signaling and asymmetric information.

Keywords: Information Asymmetries, Institutional and Cultural Economics
JEL codes: C38, D82, H4, L82, Z1

1 Introduction

In the late 19th century, a quiet rebellion was brewing in the world of art—a rebellion against the rigid constraints of academic art that dominated European culture. In France, the Academie des Beaux-Arts wielded immense influence, dictating the norms and standards of 'acceptable' art. Works that conformed to its strict guidelines on technique, subject matter, and style received prestigious salon placements and official recognition. The Königliche Akademie der Künste in Berlin and the Akademie der Bildenden Künste in Vienna served similar roles to France's Académie in the Austro-Hungarian Empire and Prussia. These institutions were instrumental in promoting neoclassical and romantic styles that embodied nationalistic sentiments and monarchical pride. The system of state-controlled academies left little room for deviation and innovation, pushing artists like Claude Monet and Pierre-Auguste Renoir to the margins. They, among others, would later be recognized as Impressionists, whose radical techniques and perspectives on light and color paved the way for modern art.

This paper looks at the European film industry and raises concerns that might mirror these historical dynamics. *Research and Markets* (2024) estimates the current market size of the European film industry at 15 billion US dollars. Unlike the US film market that is structured by large private production companies, the European film market is embedded in national and European funding institutions.[1] Central among those is the MEDIA subprogram of the European Union's "Creative Europe", which subsidizes film production and funds national film institutions. Its stated aim is to promote innovation and creativity, safeguard cultural diversity and form a shared European identity. MEDIA seeks to strike a balance between artistic innovation and commercial competitiveness in a globalized film

[1] It is difficult to estimate the exact percentage of public funding. Current statistical categories do not make a clear distinction between private and public funds. However, the private financing market for the European audiovisual sector is commonly thought to be very limited (Debande 2017).

market (European Commission 2014).

In practice, the public bodies that finance the production of films also fund the festivals that select them, the theaters that screen them and the press outlets that review them. While the influence of state-funded bodies can ensure a certain level of support and visibility of European films, centralized control might lead to conflicts of interest and even compromise creative diversity. European film productions have steadily increased during the last fifteen years. However, despite the growing number of European films, audience attendance remains flat, and with viewers still preferring US movies, each European film now attracts fewer viewers. This article provides empirical evidence for this disconnect by looking at the publication activity of the EU-funded online information portal *Cineuropa* (cineuropa.org) and how it relates to the success of European films.

Cineuropa was founded in 2002 as an open-access platform dedicated to European cinema and audiovisual productions. With daily publications of film reviews, festival reports, news from the film industry and interviews, written in four European languages (English, French, Italian, and Spanish), *Cineuropa* claims to address both industry professionals and a wider cinephile audience. At the time of writing, it counts 600,000 monthly visitors and 65,000 subscribers, making it a leading source on European cinema. Over the last 20 years, the website has published more than 5000 film reviews. It offers training programs for emerging film critics. Its articles are cited by industry-related websites such as the EU-funded streaming platform FestivalScope, or general information sites such as Wikipedia. By all appearances, *Cineuropa*'s main focus is film criticism. A potential conflict of interest stems from its institutional structure. The outlet is funded by the same financial bodies that are funding the production of films, i.e. MEDIA and national film centers of European countries. *Cineuropa* currently receives an annual grant of 250,000-300,000 euros from the EU through the MEDIA program (Creative Europe, 2024). Its list of partners includes 14 national film centers and six others are listed as former financial backers, for a total of 20

European countries that currently support *Cineuropa* or have done so in the past.

Critics are considered 'consumer guides' whose role is to inform the public about the quality of a film. This role is important because entertainment markets are characterized by asymmetric information. Films are considered to be experience goods, i.e. goods for which the quality is uncertain prior to consumption (Reinstein and Snyder 2005, Wallentin 2016). Before choosing to watch a film, viewers have access to limited information that is generated and filtered by production companies, such as promotional visuals, the synopsis and a list of collaborators. Asymmetric information can create markets for 'lemons' (Akerlof 1970). In these markets, 'lemons' are low-quality products that are presented as high-quality by dishonest sellers. Akerlof shows that, if no action is taken, sellers of 'lemons' will eventually crowd out high-quality vendors. Independent bodies such as critics are meant to guard against 'lemons' by informing consumers of the quality of a film. Conversely, when critics have incentives to side with producers, the market may be saturated by misleading signals. This can hurt critics' reputational capital as providers of reliable information and prevent consumers from making informed decisions.

In addition, the European film market displays winner-take-all dynamics typical of art and entertainment industries (see Section 2.2). A very large number of players in the film industry compete for a very limited number of rewards—be they of commercial success or artistic merit. Critical opinions become part of these dynamics. Positive reviews are expected to boost the box office performance of films with limited advertisement budget or of 'arthouse' films, which aspire to new artistic expression rather than commercial gains. Conversely, negative reviews are expected to damage big productions, regardless of their generous advertisement strategies. These assumptions are the starting point of several studies whose goal is to measure the impact critical reviews have on a film's commercial success (Eliashberg and Shugan 1997, Reinstein and Snyder 2005, Hur et al. 2016, Basuroy et al. 2020).

We build on this research by quantifying the impact of film criticism on the commercial success of European films. To the best of our knowledge, few studies focus on the film industry of European countries and predictive studies on commercial success have never considered the European film market as a whole (see Section 2.1). To identify critical reception, we use the SiEBERT (Liu et al. 2019, Hartmann et al. 2023) algorithm and attribute a sentiment score to each of the over 10,000 reviews published in *Cineuropa* and three major industry journals, *Hollywood Reporter*, *Variety* and *Screen Daily* (Section 3.1). In Section 3.2 we develop a method to estimate the commercial success of films that is based on the categories proposed by the European Audiovisual Observatory (EAO), an organization part of the Council of Europe. We attribute a commercial success class to each film in our sample of over 4,000 European films. The majority of films fall into the least successful class. We then train a XGBoost decision tree model (Chen and Guestrin 2016) to predict the commercial success of films based on distribution, festival participation, and audience ratings (Section 3.3). To handle class imbalances, we use a combination of synthetic minority oversampling (SMOTE) and Tomek Links for majority undersampling for our training data set. We add review sentiments to our model in Section 4 and use SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) to assess feature impact. We find that *Cineuropa* reviews push the model towards false positive predictions in all exercises (see Section 4.1) and interpret this as a form of false signaling. In Section 4.2 we use the same metrics but only for the most and the least successful classes. Lastly, we analyze how different market conditions affect the model outcome (Section 4.3).

Given MEDIA's focus on balancing artistic innovation with market competitiveness, a distinction can be made between commercial and artistic success. Our approach measures commercial success by focusing on admissions. However we can identify artistic success in three ways in our data. First, a major A-list festival participation can be considered an indicator of artistic success. Second, admissions can also be used as a marker of artistic success since there are audiences, albeit smaller ones, for films with artistic merit. Third,

it can be assumed that critics review a film based on its artistic potential, which may then serve readers to decide whether to watch a film or not. Therefore, even if we primarily focus on commercial success, the more qualitative variables in our data set are also artistically revealing and confirm some of the observations made in the economic analysis (see also Appendix E.3).

Film critics contribute to the public debate by rewarding certain films and questioning others. When the overwhelming majority of published reviews are positive, critics lose their capacity to provoke meaningful discussions, and change established perspectives on artistic practices. In 19th-century Europe, the rise of nationalism and imperial expansion led to the establishment of art academies, which promoted state-sanctioned artistic values that suppressed creative freedom. Today, public funding bodies and media outlets may similarly crowd out artistic diversity and consumers by privileging content that favors institutional priorities. Modern regulations on "Unfair Commercial Practice" (EU Directive 2024/825, Council of the European Union 2024) provide safeguards against such practices that did not exist in the past. While our study sheds light on problematic aspects of European cultural policy, we hope that an open discussion of the potential problems in public institutions will help to improve the future of European cinema, be it in the films produced or their critical reception.

2 The Relationship between Critique and Commercial Success

2.1 Film Criticism and Box Office Performance

Marketing scholars have previously attempted to measure the statistical significance of film reviews for box office performance. While most studies generally confirm a positive relation between reviews and box office, their findings vary significantly. In their seminal paper Eliashberg and Shugan (1997) propose two possible roles for film reviews: critics can influence the choice of consumers (*influence effect*) or can predict the commercial success of films (*prediction effect*). To study their potential influence, dynamic market data is necessary, such as weekly box office figures and exact publication dates for reviews, while the *prediction effect* can be calculated *ex post*, on the basis of a film's cumulative box office performance. Due to a wider availability of data and the size of its box office figures, much of the existing research focuses on the US market. The first of those studies ran regression models with film reviews being the independent, and box office performance (weekly or cumulative) the dependent variables. Eliashberg and Shugan (1997) find that, while film reviews are statistically insignificant for early box office revenues (*influence effect*), they are positively related to late and cumulative box office figures. According to these findings, critics can predict the response of the audience to a specific film, but the publication of their reviews has no impact on the audience's choice of watching a film or not. Reinstein and Snyder (2005) confirm the above findings about the *prediction effect*, but also find that reviews influence the box office during the opening weeks for certain film genres. On the other hand, Basuroy et al. (2003) find both a *prediction effect* and an *influence effect* for the first 8 weeks of a film's theatrical run. Separating the negative reviews from the positive, their analysis also finds that negative reviews have a stronger influence effect, suggesting the existence of a "negativity bias", i.e.

that negative opinions matter more than positive ones. Given the lack of dynamic market data for European films, the findings in our study only refer to the *prediction effect*.

The above studies take into account variables other than film reviews, such as the star power (the casting of famous actors and actresses), the franchise effect (for sequels), the genre (e.g. action, horror, etc.), and the previous work of directors or studios (awards and reputation). The only variable that consistently proves to have the strongest impact on box office performance is the total number of theaters the film screened during its opening week (number of screens), although some studies also find that a film's production budget is an important feature (Kim et al. 2013). Yet other studies propose that it is the total number of reviews written for a certain film that matters, regardless of their sentiment (Gemser et al. 2007). Less research has focused on non-US markets. Of these, Wallentin (2016) focusing on Sweden, and Gemser et al. (2007) focusing on the Netherlands, confirm a positive relation between film criticism and box office performance. To the best of our knowledge, a comprehensive study on the entire European film market does not exist.

While early research was based on reviews published in print, or on the opinions critics expressed in TV shows, later studies factored in the proliferation of non-expert online reviewing (such as audience scores). This was found to be statistically significant for box office performance (Kim et al. 2013). However Basuroy et al. (2020) find that critical opinions have a greater impact on predicting box office performance than audience scores, and that expert reviews tend to be consistently more negative. In our study, while we confirm consistently more negative reviews of industry experts (excluding *Cineuropa*), we find that audience opinions outweigh those of the press in predicting the commercial success of European films (see Section 3).

Recent research on predicting box office performance is increasingly using machine learning algorithms and classification models instead of regression analysis (Delen et al. 2007, Kim et al. 2015). Different models have been tested to predict the commercial success of a

film, such as Decision and Gradient Boosted Trees (Abidi et al. 2020, Ahmed et al. 2020), Random Forests (e Souza et al. 2023, Lash and Zhao 2016), Artificial Neural Networks (Sharda and Delen 2006) and Support Vector Machines (Du et al. 2014). Since this research is positioned within marketing studies, it prioritizes the ability to forecast a film's commercial success before its release, so that the studios can make changes in case the forecast predicts a 'flop' (Lash and Zhao 2016). As such, few prediction models include film reviews in their variables since those are published after a film's release. Some of the studies using machine learning tools turn box office prediction into a classification problem, either binary (separating 'hits' from 'flops') or multi-class (e.g. six classes or more). The thresholds that separate success from failure are set by the authors, e.g. by consulting industry experts (Lee et al. 2018) or based on personal estimations (Sharda and Delen 2006). These forecasting models confirm some of the previous findings, identifying the number of screens and the production budget as the main features contributing to a film's commercial success.

While the use of machine learning algorithms has undoubtedly increased the statistical accuracy of the models, it has sacrificed the explainability that comes with more traditional statistical methods. The preferred method of most papers using machine learning tools is to test different models and choose the one having the highest accuracy score. Our study instead uses a single model, which is known to perform well in classification tasks, and explains the ways in which each input feature contributes to the model's output. Explainability has become a key interest for machine learning and data analysis for reasons of transparency (Ramon et al. 2020). We hope to balance out any loss in accuracy that might result from not testing different models by a thorough explanation of the two models we use, SiEBERT (Hartmann et al. 2023) for sentiment analysis and XGBoost (Chen and Guestrin 2016) for the classification task (see Section 4).

2.2 Asymmetric Information and Winner-Take-All Markets

Entertainment and sports are often described as the best examples of winner-take-all markets. These are markets where very few competitors receive the majority of the available gains (Cook and Frank 1995, Elberse 2013). Commercial success (e.g. increased ticket sales) and artistic recognition (e.g. selection at a major film festival) tend to be two distinct criteria used to define a film's success. Regardless of the criteria used, only a very small number of films can be considered successful. On the one hand, the box office returns for European films are significantly lower than those for American films even in European cinemas. On the other, the acceptance rate for major festivals such as Cannes is around 2%. This means that an overwhelming majority of European films falls short of producing financial or artistic gains. The European film industry responds to this fact by investing in the promotion of European films including publicly funded journalism (European Commission 2020). To the best of our knowledge, the impact of this information in contributing to the commercial or artistic success of European films, has not been studied.

Akerlof (1970) introduced the term "Market for Lemons" to describe a market where the seller holds significantly more information about the quality of the product (e.g. its flaws and risks) than the buyer. In his model, Akerlof shows that low-quality vendors drive out high-quality vendors. Buyers eventually realize that they acquired low-quality products ('lemons'), which leads to distrust and potential market failure. To counter this information asymmetry and guard against 'lemons', laws, experts and independent bodies are needed to issue positive or negative signals about the quality of a product. Akerlof's model was first applied to the selling of used cars but other studies extended his idea (Heinkel 1981). Later research also points to conflicts of interest that can compromise defence mechanisms against asymmetric information. This takes place when sellers are controlling the access to information, or when independent bodies have financial incentives to side with sellers.

Most relevant to our study is research that shows how information asymmetry and institutional or personal interest work against consumers. Studies in different fields have shown that sellers can exert tight control over what information is published and where, which may lead to misleading signals. This makes it difficult for consumers to take informed decisions (Light and Lexchin 2021, Bridges and Vásquez 2016). Similar concerns have been expressed about potential bias in independent bodies whose role is to warn consumers of potential risks (Wiggins and Metrick 2019). As mentioned in the introduction, the European film market is characterized by asymmetric information. Positive or negative signals by intermediary bodies (such as critics and festivals) are the only available means to inform the viewers of the quality of a film (other than word-of-mouth). In a functioning market, these intermediaries can guard against 'lemons'.

Of course, films as products are very different from the used cars found in Akerlof's model, in that their perceived quality may be influenced by personal preferences and 'objective' criteria for what distinguishes a good from a bad film are more open to debate. We understand 'lemons' to be films that are not artistically or commercially promising, but are presented as such by their makers. If critics and festivals have reasons to side with the producers and distributors, then the market may overflow with positive feedback, signalling films as artistic breakthroughs or 'must-sees' regardless of their quality. To the best of our knowledge, Ravid et al.'s study (2006) of critical reviews concerning 172 films released between 1991 and 1993 in the US is the only study that has attempted to model bias in film criticism, concluding that there is a bias by some US film critics towards specific US production studios. On the other hand, bias in film criticism has never been investigated for the European film market.

3 Data Overview and Empirical Methodology

3.1 Sentiment Analysis of Film Reviews

We used various web-scraping tools to retrieve 5760 film reviews from cineuropa.com between 2003 and 2022. Reviews of the same films were retrieved from three industry journals, *Variety*, *Hollywood Reporter*, and *Screen Daily*. The choice of these outlets was driven by their importance for the film industry. Additionally, we obtained IMDb ratings and Rotten Tomatoes' audience scores. An overview of our data set is provided in Table 1. While 912 reviews are present in all four film journals, combining the industry outlets leaves us with 3014 shared reviews between *Cineuropa* and the industry press.

Table 1: European Film Review Data Sets

Medium	N.(obs)	% of Cineuropa	Mean Sent.	Std.
Cineuropa	5760	100	0.76	0.16
IMDB Audience Ratings	4185	72.66	0.65	0.07
Variety	2082	36.15	0.59	0.19
Screen Daily	2049	35.57	0.67	0.17
Hollywood Reporter	1837	31.89	0.62	0.17
Rotten Tomatoes Audience Ratings	1423	24.70	0.62	0.19

Note: Overview of film reviews from *Cineuropa*, corresponding industry reviews in *Variety*, *Hollywood Reporter*, and *Screen Daily*, along with Rotten Tomatoes and IMDb ratings. % values report on the percentage of overlapping films. Columns Mean Sent. and Std. report the average sentiment score and standard deviation of each data set.

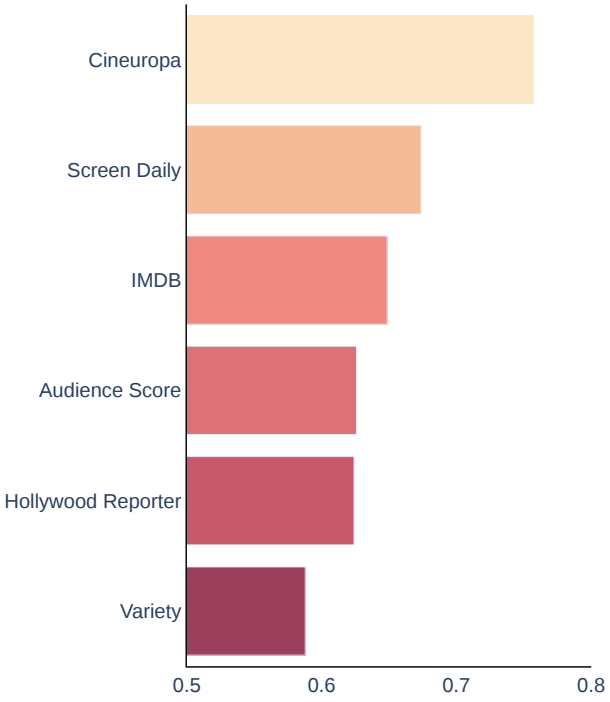
We use SiEBERT-large (Hartmann et al. 2023), a fine-tuned checkpoint of RoBERTa-large (Liu et al. 2019). SiEBERT performs sentiment analysis on different types of English text, predicting positive (1) or negative (0) sentiments. It was fine-tuned on 15 diverse data sets—including reviews and tweets. This leads to better performance on new texts compared to models trained on a single type, like the SST-2 movie reviews benchmark. Detecting sentiment in movie reviews has become a standard benchmark for evaluating the

performance of sentiment analysis models. SiEBERT was trained on several such data sets including Pang et al. (2002), based on 2000 IMDb reviews; Pang and Lee (2005), based on 10,662 Rotten Tomatoes snippets; Maas et al. (2011), based on 50,000 IMDb reviews.

One difference between these data sets and our movie reviews is that our texts tend to be longer. Transformer models like BERT (Devlin et al. 2019) use a fixed number of input tokens (512), which can limit their ability to process and analyze longer texts. Different methods have been proposed to circumvent this problem, a common solution being slicing or splitting (Dai et al. 2022). We use the latter approach and parse each text to the sentence-level. We then feed each sentence into the model and aggregate the mean score for each document. This approach assumes that a review can express both negative and positive sentiments and that the final sentiment of a review will rarely be completely negative (0) or positive (1). The results of this process are reported in Figure 1 and Table 1. Previous studies on the sentiment scoring of online movie reviews report that the absence of domain-specific dictionaries and the difficulty in mimicking the human recognition process makes sentiment analysis a complicated task (Hur et al. 2016). We consider our approach to be an improvement of previous methods. On the one hand, SiEBERT does not require a manually created domain-specific dictionary. On the other, the sentiment score for each review is expressed as a percentage of sentences recognized as positive. As such, its sentiment score is best understood by comparing it to the score of other reviews, and not to the reader's interpretation of it.

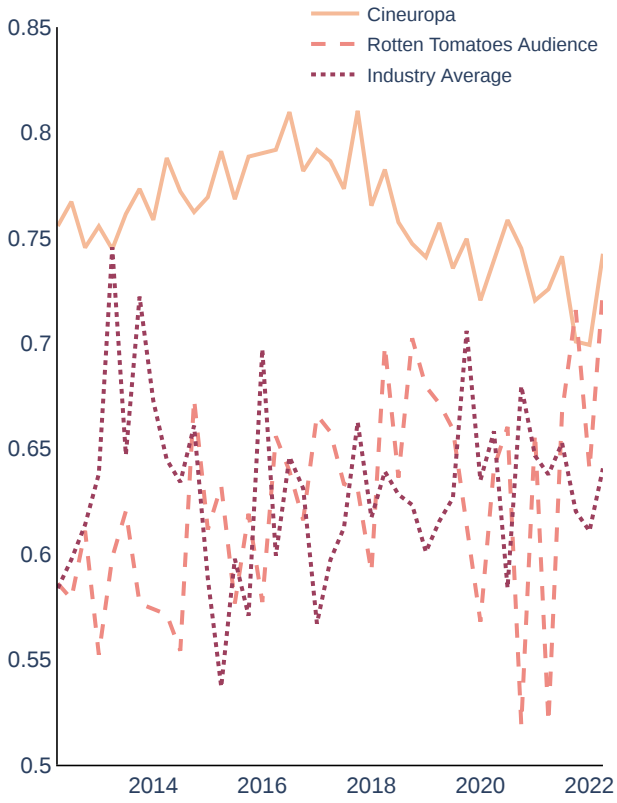
With an average sentiment score of 0.63, the industry press seems to be closer to both IMDb and Rotten Tomatoes' audience ratings than to *Cineuropa*, which has the highest average sentiment score at 0.76. These findings are similar to previous research including a 'neutral' category for reviews not expressing clear sentiment, which find that non-negative reviews account for 60 to 75% of the total number of reviews (Eliashberg and Shugan 1997), Basuroy et al. (2003). The general skewness towards positive values—all data sets have

Figure 1: Sentiment Scores of Media Outlets



(a) Average Sentiment Score

Note: Mean sentiment scores of film reviews in our data set. Sentiment scores are obtained using SiEBERT (Hartmann et al. 2023), where 0 is negative and 1 is positive. With 0.76 *Cineuropa* reviews are the most positive and *Variety* the least positive with 0.59.



(b) Average Sentiment over Time

Note: Quarterly aggregates of mean sentiment scores. Industry average is calculated as the average scores of *Variety*, *Hollywood Reporter*, and *Screen Daily*.

an average above 0.5, does not necessarily indicate positive reviews (see also Figure A.1 in the online Appendix). A sentiment score of 0.5 indicates that the review contains an equal number of positive and negative sentences. This does not mean, however, that positive and negative sentences are equally perceived by the reader. Research has shown that negative opinions often have a greater influence on impression formation and social judgement, referred to as "negativity bias" (Skowronski and Carlston 1989). Basuroy et al. (2003) use this concept to examine whether negative reviews have a greater impact on box office performance than positive ones. Similarly, a few highly negative sentences can have a significant

impact on the way a reader perceives the text, thus a review with a sentiment score above 0.5 is not always perceived as positive. A similar logic appears to underpin Rotten Tomatoes, where a film needs to pass the 60% threshold to be considered 'fresh'. A more detailed explanation on how SiEBERT assigns sentiment to individual sentences is provided in the Appendix B.3.

3.2 Estimating Commercial Success

As discussed in Section 2, existing research on the US market generally aims to help investors make informed decisions when funding new film productions. These predictive models classify the commercial success of films either based on individual profits or gross box office revenues (Lash and Zhao 2016). Because what matters for private investors is to know how profitable a specific film may be, i.e. their possible return on investment, these models tend to use variables available before the release of the films (e.g. actors, genres, MPAA rating, previous work of studios or directors, etc.). However, this approach is not suitable for the European market, as production budgets and box office revenues are rarely publicly known figures. Previous studies have used admissions (i.e. the number of paid tickets) instead of data on box office (Meiseberg and Ehrmann 2013).

The European Audiovisual Observatory (EAO) proposes admission thresholds for classifying the commercial success of European films. In 2023, the EAO published a classification method on the basis of six market clusters and four categories of commercial success (Kanzel 2023, and table E.2 in the online Appendix). However, this method has two important drawbacks. First, it does not account for variations over time. Second, the proposed thresholds are relatively high, meaning that, in practice, almost all European films would fall into the least successful class (96% in 2022).

In our study, we introduce a different method to define admission thresholds for five

commercial success categories. To address geographical inequalities, e.g. different policies of film distribution and infrastructure, that may emerge from considering European market as a single market, we follow the EAO annual reports (Kanzel 2018-2022, 2024) which divide the European countries into three market clusters. We attribute a market size (S, M, L) to each film based on its country of production or the first country in case of co-productions (see Table E.1 in the online Appendix). To estimate each film's commercial success, we need to know their market size and their number of admissions. We use the LUMIERE database (lumiere.obs.coe.int) to extract the total admissions of each film. The admissions correspond to the number of paid tickets for films with theatrical release in at least one of the European countries represented in the database.

Table 2 presents the conditions we use to classify the films in respect to their commercial success. As mentioned above, it is impossible to follow the typical method of comparing productions costs to revenues. Instead, we use the median production budget B_c and the real average ticket price P_c as a proxy (estimated per year and per market). We assume that a film has high commercial success when its admissions A_f are higher than the median production budget B_c divided by the average ticket price P_c of its market cluster. Conversely, a film has low commercial success when its admissions only reach a quarter (1/4), or half (1/2) of the estimated B_c/P_c. We use five success categories S_c (*micro, low, medium, high and very high*). The admissions thresholds are then calculated separately for each market.

We obtained the average ticket price for each European country from industry statistics published by the EAO at the occasion of the 2022 Cannes Film Festival (Kanzel and Simone 2022). The publication included ticket prices only for 2021. We use the consumer price index to estimate the ticket prices for the remaining years. We retrieved the median production budget for each market from the EAO annual reports (Kanzel 2018-2022, 2024). For the years where these figures were unavailable (2008 to 2015), we estimate the missing production budgets by assuming that they follow the yearly fluctuation of the production budget of

Table 2: Admission Thresholds for Commercial Success Categories

Condition	**Commercial Success (S_c)**
$A_f < 0.25 \times (B_c/P_c)$	$S_c = micro$
$0.25 \times (B_c/P_c) < A_f < 0.5 \times (B_c/P_c)$	$S_c = low$
$0.5 \times (B_c/P_c) < A_f < B_c/P_c$	$S_c = medium$
$B_c/P_c < A_f < 1.5 \times B_c/P_c$	$S_c = high$
$A_f > 1.5 \times (B_c/P_c)$	$S_c = very\ high$

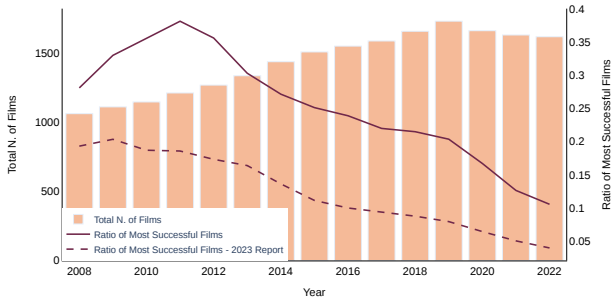
Note: The thresholds for the admissions are calculated for each market size (S, M, L) using the median production budget and the average ticket price of each year (2008-2022). A manual comparison between the assigned commercial success class and the known success of individual films seems to verify the estimated admission thresholds (see Appendix E.1). A_f: Paid ticket admissions for all years of distribution; P_c: Real average ticket prices per year, per market size; B_c: Median production budget per year, per market size; S_c: Success category: $S_c = \{micro, low, medium, high, very\ high\}$.

French films. Given the importance of the French film industry (see also Section 4), we use this method as a fallback due to the lack of historical data. The budget data for French films was retrieved from the annual reports of the Centre national du cinéma et de l'image animée (cnc.fr). We use the alternative classification method of the EAO 2023 Report throughout our study to cross-verify our findings.

We separate the least commercially successful films (class S_c calculated as *micro* or *low*) from the more successful ones (class S_c calculated as *medium*, *high*, or *very high*), attributing a new class S_b of '0' to the former and '1' to the latter (see Table 3). Figure 2 shows the ratio of most successful films in our sample (class '1') per year. The number of successful films each year is very limited. After a certain point, releasing more films does not increase success (see the marginal rate of success in Figure 2.b). This is consistent with the behavior of winner-take-all markets as described in Section 2. The ratio of successful films is even lower when the same method is applied to a data set including all European films released between 2003 and 2022 (see Figure E.1 in the online Appendix).

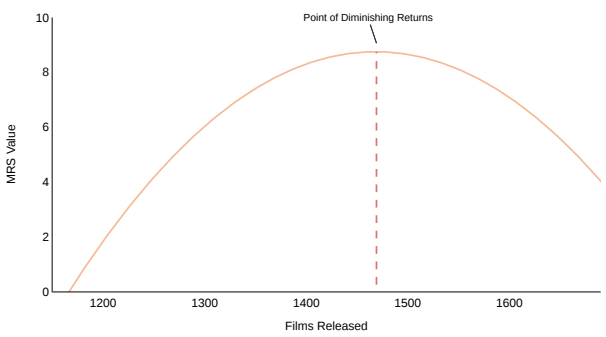
Our baseline model predicts the commercial success of films based on the number of

Figure 2: Commercial Success and Number of Film Productions



(a) Ratio of Most Successful Films in Our Sample

Note: Figure (a) shows the rolling average of the total number of European films released per year and the rolling average of the ratio of the most successful films (class '1'). Class '1' is attributed to films with admissions $A_f > 0.5 \times (B_c/P_c)$, where B_c is the mean production budget and P_c is the average ticket price per year for each market (Class '1' combines the medium, high, and very high success categories S_c, see Table 2). The mean value of the ratio of most successful films is 0.24. The dashed line shows the ratio for the most successful films classified according to the EAO 2023 Report (Kanzel, 2023).



(b) Marginal Rate of Success

Note: Figure (b) shows the marginal rate of success (MRS) for the most successful films (class '1'). A second-degree polynomial curve was calculated to fit the MRS values. For positive MRS values close to 0, a significant increase in the number of films is necessary to have more successful films. The maximum point in the curve corresponds to an MRS value of 8.75 and 1470 films produced. After that point, increasing the number of films does not result in a proportionate increase of successful films.

distribution countries, the audience score and the festival selection of each film (see Table 3 for an overview of all the variables used in the study. The *Distribution* variable represents the distribution countries from the LUMIERE database. It is equivalent to the total number of theaters the film screened, which represents the most common variable used in box office prediction models for US films (see Section 2). From IMDb (imdb.com), we use the festival run of each film to create a proxy for festival success. We attribute a value of '0', '1' or '2' to each film, where '2' corresponds to an A-list festival selection, according to the classification of festivals by the International Federation of Film Producers Association (FIAPF), '1' to any non-A-list festival selection, and '0' to films without festival circulation. From IMDb, we also retrieved the audience rating for each film.

The number of films with full information on both their critical reception and commercial

Table 3: List of Variables Used in the Analysis

Variable	**Description**	**Values**
$Fest_f$	Festival selection per film	(0, 1, 2)
Distribution	N. of distribution countries	public data (log transf.)
IMDb	Average Audience Score	(0.0 - 10.0)
S_b	Commercial Success (Binary)	(0, 1)
A_f	N. of paid admission per film	public data
P_c	Average Ticket Prices (per market/year)	public data (est.)
B_c	Median Prod. Budget (per market/year)	public data (est.)
Market size	The Market Size of the Prod. Country	(S, M, L)
Industry Avg	Average Sentiment Score - Industry	(0.0 - 1.0)
Cineuropa	Average Sentiment Score - Cineuropa	(0.0 - 1.0)

Note: The four variables in the upper part of Table 3 are included in the baseline XGBoost model. The four variables in the middle part are used to estimate the commercial success categories. The lower part shows the sentiment score variables, where *Industry Avg* is the mean of the sentiment scores attributed to the reviews of three industry outlets. *Distribution* was log transformed. Sources: IMDb, LUMIERE, EAO and own calculations.

releases vary for each year. The total number of critically reviewed films without missing values for all market variables amounts to 4181 films or 19% of the estimated European films released between 2008 and 2022 (see Table E.3 in the online Appendix). Our sample size is on the larger end compared to those typically used in similar research. Most studies use 300 to 600 films (Basuroy et al. 2020) and few use more than a couple of thousand observations (e Souza et al. 2023).

In the next section we train a XGBoost model on our data set. If the accuracy of the model is high, we can make assumptions about the meaning of different variables contributing to the model output because the model has learned patterns from the data from which to make reliable predictions. We separate 25% of the data for testing and 75% for training. The train-test logic involves splitting the data set, where the model learns from labeled examples and is evaluated on unseen data.

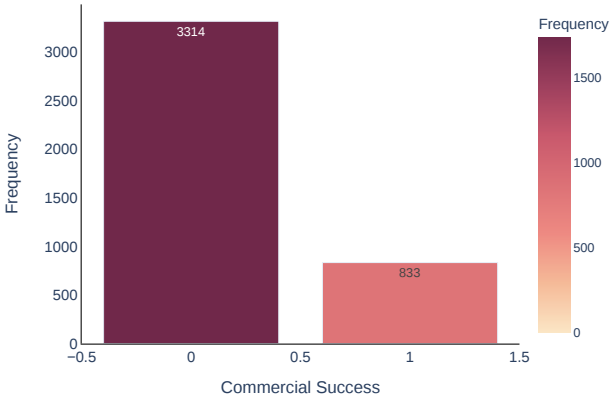
3.3 Binary Classification of Imbalanced Data

A majority of films in our data set fall into the category of least successful films (see Figure 3). Imbalances in class distribution within data sets can degrade machine learning performance, as models may become biased toward the dominant class (He and Garcia 2009). Several solutions exist for handling imbalanced data sets (for a survey, see Johnson and Khoshgoftaar 2019). We use SMOTETomek to solve the class imbalances in our data set. SMOTETomek is a hybrid method that combines the SMOTE (Synthetic Minority Over-sampling Technique) and Tomek Links techniques. SMOTE works by creating synthetic samples from the minority class by interpolating between existing samples (Chawla et al. 2002). Tomek Links identifies pairs of very close instances, but of opposite classes (Tomek 1976). Removing the majority class instances of these pairs helps in cleaning overlapping data points between classes, thus making the decision boundary between classes clearer. Introduced first by Batista et al. (2003), SMOTETomek combines minority oversampling and majority undersampling (see Appendix C.1 for a detailed explanation). Figure 3 gives an overview of the balanced data set. Note that the SMOTETomek technique was applied only to the training fold.

We next structure the feature matrix for predictive modeling. We define the feature matrix $X = [x_1, x_2, x_3, \ldots, x_n]$ to include IMDb audience ratings, the log-transformed *Distribution* variable indicating the number of countries in which a film was shown and the $Fest_f$ variable representing the categorical data on festival participation. Our baseline XGBoost model (Chen and Guestrin 2016) is built on the resulting dataset (see also Table 3 for an overview of the variables used).

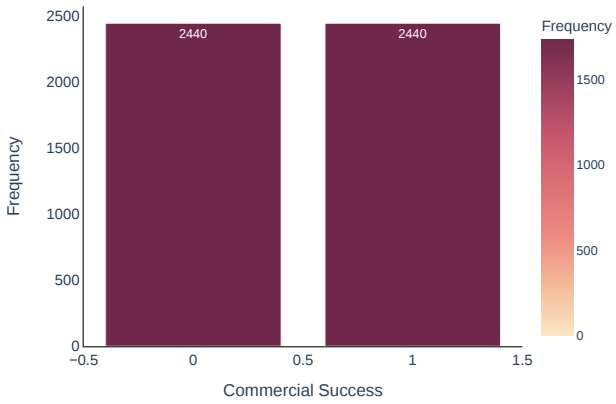
XGBoost uses a gradient boosting framework for decision trees. Gradient boosting is an ensemble technique that builds models sequentially, where each new model progressively reduces the errors of the previous models. Models are added until no further improvements can be made. It combines multiple weak or average predictors to build a stronger predictor.

Figure 3: Imbalanced and Balanced Data Set



(a) Initial Distribution of Commercial Success Categories

Note: The original data set shows a significant imbalance in the distribution of the commercial success categories. The majority of data points are classified in the '0' (least successful) class with 3314 instances, followed by much smaller counts in *medium* to *very high* success (class '1').



(b) Distribution of Commercial Success Categories after SMOTETomek

Note: The data set has been balanced using the SMOTETomek technique, resulting in a more uniform distribution of commercial success categories. Each category now has the same number of instances.

This method is useful for imbalanced data because it focuses on correcting the mistakes of previous trees in the sequence. Each subsequent tree prioritizes harder-to-classify instances, e.g. the minority classes in imbalanced data. In binary classification, the objective function of XGBoost can be expressed as:

$$\text{obj}(\theta) = \sum_{i=1}^{n} \mathbf{L}(y_i, F(\mathbf{x}_i)) + \sum_{k=1}^{K} \mathbf{\Omega}(f_k), \tag{1}$$

where $\mathbf{L}(y_i, F(\mathbf{x}_i))$ is the binary log loss, which measures the discrepancy between the actual labels y_i and the predicted probabilities $p_i = \sigma(F(\mathbf{x}_i))$. $F(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i, \theta_k)$ represents the sum of the outputs of all the decision tree models up to the current iterations, and $\sigma(\cdot)$ is the *sigmoid* function, which converts these outputs into probabilities. The regularization term $\mathbf{\Omega}(f_k)$ penalizes the complexity of the model. This term includes components for the number of leaves and the weights of the leaves within each tree and controls the model complexity through L1/L2 regularization (see Appendix C.2 for a detailed explanation on

gradient boosting and the hyper-parameter settings).

Table 4: Classification Report for Baseline XGBoost

Class	Precision	Recall	F1-Score	Support
0	0.92	0.83	0.87	817
1	0.54	0.74	0.62	220
Accuracy			0.81	1037
Macro Avg	0.73	0.78	0.75	1037
Weighted Avg	0.84	0.81	0.82	1037

Note: Performance metrics of a classification model for predicting commercial success of European films from audience scores, distribution and festival attendance. The metrics include precision, recall, F1-score, and support numbers for each class. Class '0', representing the least successful films, displays the strongest results, while more successful films, class '1', show a markedly weaker performance.

Table 4 provides an overview of the performance metrics of the baseline classification model (the learning curves are reported in Figure A.4 of the Appendix). The majority class '0', the least successful films, shows the strongest performance with a precision of 0.92, recall of 0.83, and an F1-score of 0.87. In contrast, more successful films (class '1') demonstrate significantly weaker performance, with a precision of 0.62. The overall accuracy of the model is 81%. The macro average of 75%, while still acceptable for making generalizations, indicates the model's limited effectiveness in predicting more successful films. Despite applying class balancing techniques, predicting the minority class thus remains difficult. Previous studies report similar results for their best performing binary classification models. Lash and Zhao (2016) and e Souza et al. (2023) report a F1 score of 0.68 and 0.96 respectively for a Random Forest model, while Ahmed et al. (2020) report an F1 score of 0.82 for a hybrid model including predictions by XGBoost trees and Random Forests. We further test the robustness of our model by using the alternative classification for commercial success proposed by the EAO 2023 Report (see Table E.2 in the online Appendix) and find that this method leads to a significantly lower prediction accuracy (see Table C.1 in the online Appendix).

Interpreting the output of machine learning models is difficult, particularly with complex

models where accuracy often comes at the expense of transparency (Hassija et al. 2023).[2] A common approach for linear models is to use nested model comparisons or incremental F-tests to assess the impact of additional variables on the model's explanatory power. However, these traditional statistical methods do not work well for non-linear models because they lack the straightforward interpretability of linear coefficients and hypothesis testing framework.

More recently, interpretability techniques like SHAP (SHapley Additive exPlanations), have been developed to provide insight into how different features contribute to individual predictions (Lundberg and Lee 2017). SHAP assigns each feature an importance value for a particular prediction. SHAP values are grounded in cooperative game theory (Shapley 1951). Shapley values attribute a value of contribution to each player (feature) in a coalition (model prediction), based on the concept of fairly distributing both gains and costs to participants according to their contributions. The technical basis of SHAP involves the decomposition of a prediction into the contributions of each feature. This method calculates the marginal contribution of each feature by considering all possible permutations of features and averaging their contributions across these permutations. The sum of all feature contributions equals the difference between the expected output and the actual output of the model. A more technical description of SHAP is provided in Appendix D. In the next Section, SHAP is used to examine the impact of our film review sentiment indicators in classifying success.

[2] Studies have shown that in-built tools, which compute importance as the mean and standard deviation of accumulation of the impurity decrease within each tree, are not always reliable in explaining feature importance and XGBoost's 'black box' functions (Adler and Painsky 2022).

4 Empirical Findings

4.1 The Impact of Film Criticism on Commercial Success

In a first exercise, we compare models that include different combinations of sentiment scores—from *Cineuropa*, *Industry Average*, or both—to understand which sources of sentiment data might alter the predictive accuracy. It is important to note that predicting success from review sentiments does not mean identifying 'blockbusters' or 'crowd-pleasers', but that critical and audience reception share some common ground. In other words, if critics exclusively like films that audiences do not, the relevance of film journalism for the wider public is called into question.

We focus our analysis on the false positive ratio of each model, i.e. the proportion of negative instances that are incorrectly defined as positive. The false positive ratios reveal how positive sentiments from reviews can skew the predictive models toward inaccurately labeling films as likely to be successful when they might not be. The result of this exercise is reported in Table 5. Note that the data set used for this and the preceding exercises counts 2309 observations as we do not have observations for all industry press sentiment scores. We observe that integrating sentiment scores from both *Cineuropa* and the industry averages does not increase the predictive capacities of our baseline model. In fact, all models share the same validation metrics including the false positive ratio of 17.29%. Using the classification method of the EAO 2023 Report, we find more varying FP ratios, with the combined model having the lowest score.

While adding or removing features from the model may be helpful in observing changes in predictive performance, it only provides limited insights into the specific contributions of each feature. It is possible that multiple features interact in a way that either increases or

Table 5: Classification Report for Sentiment Models

Model	Precision	Recall	F1-Score	FP Ratio	FP Ratio EAO	Support
Baseline	0.72	0.76	0.73	17.29	18.32	578
Cineuropa	0.72	0.76	0.73	17.29	16.79	578
Industry Avg	0.72	0.76	0.73	17.29	16.6	578
All	0.72	0.76	0.73	17.29	15.56	578

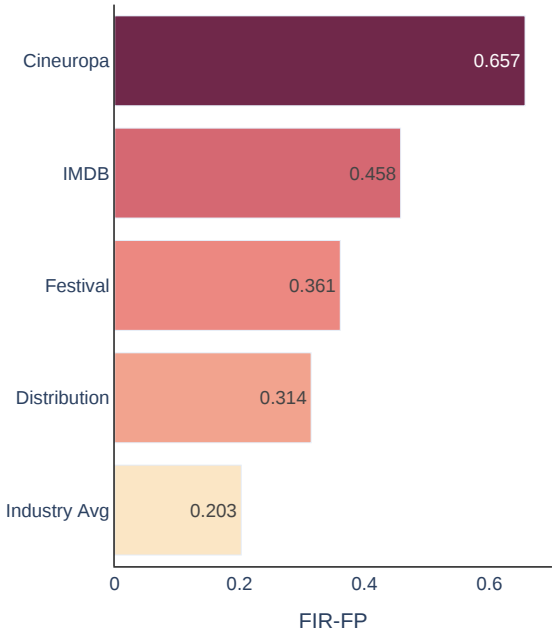
Note: Macro Precision, recall, F1-score, false positive (FP) ratios for our success metric and the EAO method, and the number of observations (Support) used in the test set for each model. The models include the baseline (no sentiment scores), Cineuropa (incorporating sentiment scores from Cineuropa), Industry Avg (using industry average sentiment scores), and All (combining both Cineuropa and industry scores). The lowest false positive scores are reached by the combined model.

decreases the false positive rate, making it difficult to attribute changes to specific features by ablation alone. To gain a more detailed understanding of the importance of each feature, we use SHAP values. As mentioned in the last section and in Appendix D, SHAP provides a method for breaking down the contribution of each feature to a prediction.

We observe that *Distribution* is by far the most influential feature, confirming the relationships we also observed in the correlation matrix (Figure A.3 of the Appendix). This is confirmed for the model using the classification method of the EAO 2023 Report (see Figure A.6 in the online Appendix). This observation is in line with Eliashberg and Shugan (1997), Reinstein and Snyder (2005) and Kim et al. (2013). Unlike previous studies (Basuroy et al. 2020), we find that audience opinions represented in the *IMDb* variable is the second most important feature, more important than critical reviews. The opinion of the press, be it *Cineuropa* or the industry press, appears to have the least impact on the model output, although *Cineuropa* scores higher than the $Fest_f$ variable when using the classification of the EAO 2023 Report. For both *Distribution* and *IMBb* the SHAP values show that higher distribution and audience ratings increase the likelihood of success. The SHAP values of *Cineuropa* show that more positive reviews decrease the likelihood of success. Most of the SHAP values of the industry press are centered around 0 with some outlier values on both

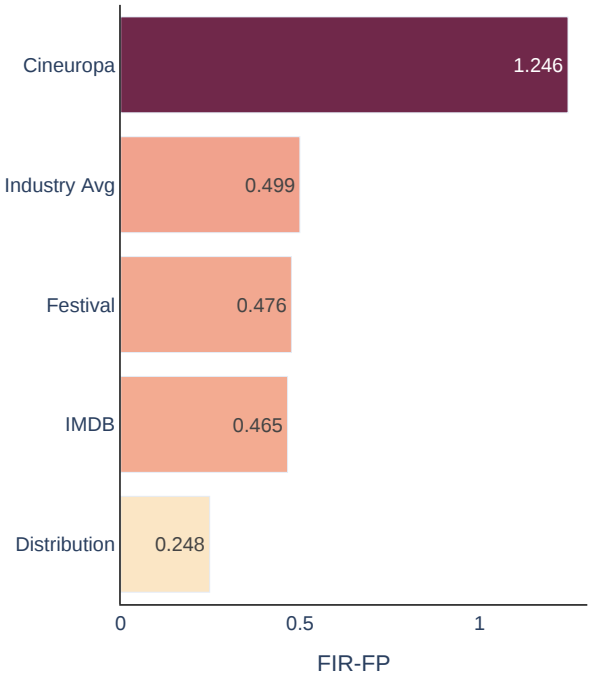
sides of the spectrum. Lastly, an interesting observation can be made from the SHAP values of the festival $Fest_f$ variable, where not being selected at an A festival appears to push the model towards higher output values (i.e. success). This may be explained by the presence of blockbuster films which circumvent the festival circuit, an art vs. entertainment conflict we address in Appendix E.3. Examples of films and their SHAP decomposition are illustrated in Figure A.8 in the online Appendix.

Figure 4: Feature Impact Analysis for False Positives



(a) Feature Impact Ratio for False Positives (FIR-FP) Baseline Model

Note: The FIR-FP metric combines the tendency of each feature to push predictions toward the positive class, its contribution to false positives, and its weighted SHAP value impact. Higher FIR-FP values indicate that a feature disproportionately increases the likelihood of positive predictions.



(b) FIR-FP EAO 2023 Report

Note: FIR-FP derived using a model with a different success classification method from the EAO 2023 Report of the European Audiovisual Observatory (Kanzel 2023). The model using the EAO method has more false positives, which may be the reason for the higher FIR-FP values.

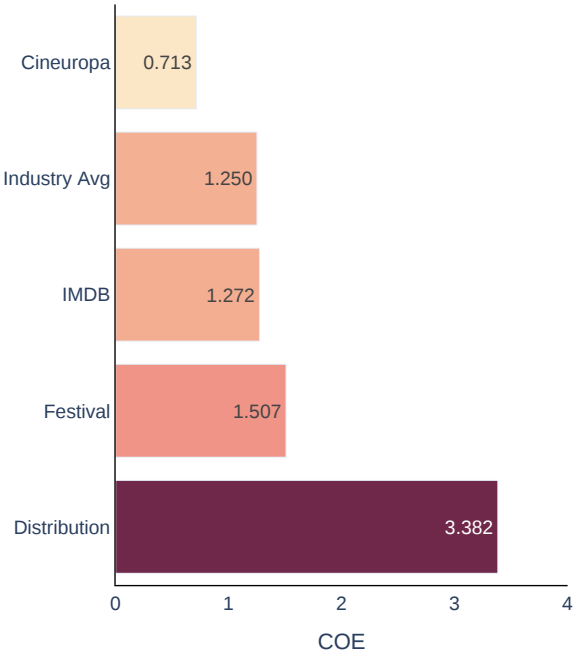
In this paper, we are most interested in understanding the specific contributions of individual features to the false positive rate of our predictive model. We use SHAP to estimate a Feature Impact Ratio for False Positives (FIR-FP). Let ϕ_0 denote the expected value of the SHAP explanation, which is the SHAP base value or the mean prediction for

a sample if no feature information is available. The Positive Push Ratio PPR_i represents how much feature i pushes the model towards positive predictions. It is calculated as the ratio of the number of positive SHAP values to negative SHAP values for feature i across all instances: $\text{PPR}_i = \frac{\sum_{j \in \mathcal{D}} \mathbf{1}(\phi_{i,j} > \phi_0)}{\sum_{j \in \mathcal{D}} \mathbf{1}(\phi_{i,j} < \phi_0)}$, where $\phi_{i,j}$ is the SHAP value for feature i in instance j, ϕ_0 is the expected value, and $\mathcal{D}$ represents the set of all instances. The False Positive Push Ratio FPPR_i represents the ratio of the number of positive SHAP values in false positive instances to the sum of positive and negative SHAP values for feature i within false positive instances: $\text{FPPR}_i = \frac{\sum_{j \in \mathcal{FP}} \mathbf{1}(\phi_{i,j} > \phi_0)}{\sum_{j \in \mathcal{FP}} (\mathbf{1}(\phi_{i,j} > \phi_0) + \mathbf{1}(\phi_{i,j} < \phi_0))}$. Here, $\mathcal{FP}$ denotes the set of false positive instances, where the model predicted a positive outcome ($p_i = 1$) but our classification labeled it as '0' ($y_i = 0$). Finally, the Feature Impact Ratio for False Positives (FIR-FP) is defined as the product of PPR_i and FPPR_i, to take into account the general push towards positive predictions and the specific push within false positives: $\text{FIR-FP}_i = \text{PPR}_i \times \text{FPPR}_i$.

Figure 4 reports the results of the FIR-FPs. The FIR-FP scores for the features in our baseline model range from 0.2 to 0.66. This means that none of these features strongly contribute to the false positive rate. A score below 1 suggests a minimal impact on false positives, while slightly higher scores, such as for *IMDb* and *Cineuropa* indicate a somewhat greater, but still moderate, impact. Nevertheless, we note that *Cineuropa*, even though it does not rank high in its overall impact, disproportionately contributes to false positives. On the other hand, while *Distribution* has the highest overall impact of the model output, it has a much lower FIR-FP. For the model using the classification method of the EAO 2023 Report, *Cineuropa* stands out with a significantly higher score of 1.28 and thus has a much greater influence on pushing the model towards false positive predictions compared to the other features for this model. We confirm our findings by calculating the FIR-FPs of the model including *Industry Average* but not *Cineuropa*, and observe that the FIR-FPs of all features are below 1 in both the baseline and the EAO 2023 Report model.

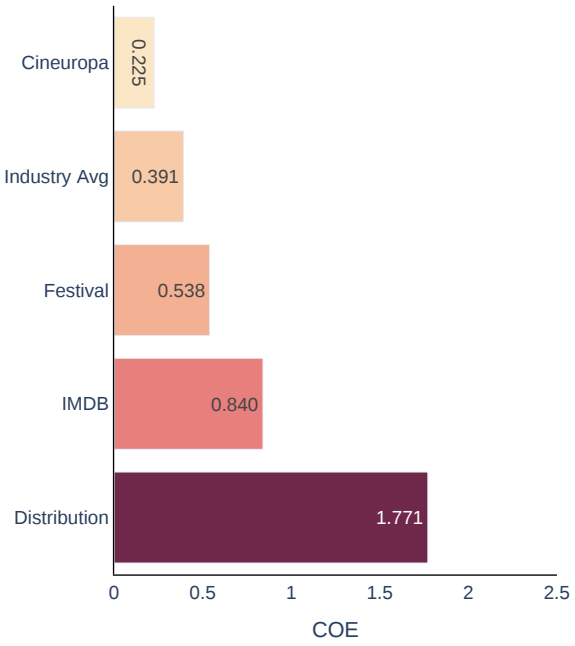
We next calculate a Crowding Out Effect, defined as $\text{COE}_i = \frac{\text{TP}_i}{\text{FP}_i} \times \frac{1}{\text{PPR}_i}$, where TP_i is

Figure 5: Crowding Out Effects



(a) Crowding Out Effect Baseline Model

Note: COE values indicating how well each feature balances true positives against false positives. A COE less than 1 means the feature contributes more to false positives than to true positives, after accounting for its PPR.



(b) Crowding Out Effect EAO 2023 Report

Note: COE derived using a model with a different success classification method from a report of the European Audiovisual Observatory (Kanzel 2022).

the number of true positives associated with feature i, FP_i is the number of false positives associated with feature i, and PPR_i is the Positive Push Ratio for feature i. A COE value greater than 1 indicates that the feature contributes more true positives than false positives, while a COE value less than 1 means the feature generates more false positives than true positives. The results of the analysis are reported in Figure 5. The COE values in both the baseline and the EAO 2023 Report models, show that *Distribution* consistently has the highest values, meaning it is effective at identifying successes without overwhelming the model with false positives. *Cineuropa* has the lowest COE among the features, contributing more false positives than true positives to the model. However, in the EAO 2023 Report model, we also observe a COE below 1 for other variables, albeit less markedly so.

4.2 Winner-Take-All

We next estimate the parameters of a model that focuses exclusively on films that are either clear commercial successes or clear commercial failures according to the taxonomies described in Section 3 (with $S_c = micro$ or S_c=*very high* according to Table 2). The reason for this adjustment stems from the observation that, as we discussed in Section 2.2 the film industry behaves as a winner-take-all market, where a lot of films compete for very few spots of success. The results of the winner-take-all model are reported in Table 6. The model shows better accuracy scores with an overall F1-score of 0.84 for the combined model, compared to 0.73 for the model that included more granular thresholds of commercial success. This improvement is further reflected in the lower false positive ratios, which were reduced by around 10% for all models.

Table 6: Classification Report for Winner-Take-All Models

Model	Precision	Recall	F1-Score	FP Ratio	FP Ratio EAO	Support
Baseline	0.78	0.90	0.82	7.36	8.78	439
Cineuropa	0.79	0.90	0.83	6.60	8.55	439
Industry Avg	0.78	0.90	0.82	7.87	8.31	439
All	0.80	0.90	0.84	7.36	8.78	439

Note: Macro Precision, recall, F1-score, false positive (FP) ratios for our success metric and the EAO method, and the number of observations (Support) for each model. The models include the baseline (no sentiment scores), Cineuropa (incorporating sentiment scores from Cineuropa), Industry Avg (using industry average sentiment scores), and All (combining both Cineuropa and industry scores). The highest accuracy is achieved by the combined model.

We repeat the SHAP analysis for the winner-take-all model. *Distribution* consistently stands out as a reliable feature, with the lowest FIR-FP values (0.18 in the baseline model and 0.08 in the EAO 2023 Report model), indicating it contributes minimally to false positives. Additionally, its COE is significantly high (7.3 in the baseline and 4.6 in the EAO 2023 Report model), reinforcing its effectiveness in correctly distinguishing true commercial successes from failures. The model also confirms *Cineuropa* as a significant contributor to false positives.

In the winner-take-all scenario, its FIR-FPs are again the highest among the features (0.74 in the baseline model and 0.68 in the EAO 2023 Report). The COE values for *Cineuropa* are 1 and 0.42 for the baseline model and the EAO Report, respectively, indicating that its reviews inflate the success chances of films.

The winner-take-all model's focus of analyzing clear successes or failures seems to better capture the film industry's competitive landscape and confirms the findings of previous FIR-FP and COE analyses. *Distribution* proves to be a robust predictor of true success, while *Cineuropa* continues to skew predictions by overestimating success.

4.3 Centralized and Decentralized Competition

The European film market may be a winner-take-all market, but it is not a competitive market in many ways. As discussed in Section 2, funding is provided by local, national and supra-national government bodies, which may influence the production and distribution of films. Countries like France, which have established cultural policies and significant subsidies, often have a stronger presence in the market. For the films in our sample, we observe that over 30% of films are either produced or co-produced by France and that almost 50% of films are distributed in France. This makes France the largest production and sales market in Europe. We compute dummy *Competition* variables that look at both the production and distribution involvement of countries and include them in our baseline model. A country is considered dominant if it appears frequently in both production and distribution lists beyond a specified threshold and is marked as '0'. We set the threshold at 20% which only includes France (a threshold below 15% would also include Germany).

Isolating France is in line with the EAO annual reports where the studies show results first for the entire data set and then for a subset that excludes French films. In a market where one player has a substantial advantage, the competitive dynamics are skewed, which

makes it more difficult for other countries to exert influence or challenge the dominant player. We therefore consider the French market to represent a 'low' ('0') competition environment. Alternatively, high competition ('2') is assigned when a country has higher distribution frequency compared to production, and does not frequently co-produce with France (<10%). All other scenarios default to a medium competition level ('1'). This leaves us with 33% (1359) of films categorized as '0', 53% of films categorized as '1' and 13% of films categorized as '2'. The classification report of the baseline model including *Competition* is reported in Table 7.

Table 7: Classification Report for Winner-Take-All Model with Competition Dummies

Model	Precision	Recall	F1-Score	FP Ratio	FP Ratio EAO	Support
Baseline	0.79	0.90	0.83	6.6	9.56	439
Cineuropa	0.80	0.90	0.84	6.35	–	439
Industry Avg	0.80	0.91	0.84	6.35	–	439
All	0.79	0.93	0.93	6.6	–	439

Note: Macro Precision, recall, F1-score, false positive (FP) ratios, and the number of observations (Support) for each model. The models include the baseline (no sentiment scores), Cineuropa (incorporating sentiment scores from Cineuropa), Industry Avg (using industry average sentiment scores), and All (combining both Cineuropa and industry scores). Because of extremely low SHAP values for the EAO 2023 Report method, we only ran the baseline model for this exercise.

We repeat the SHAP analysis for this model and confirm our previous observation that *Distribution* remains the most important feature in the model. However, the medium competition dummy appears to have a generally higher impact on the predictive outcome than the review sentiment scores and IMDb, making it the most important predictor after *Distribution* and $Fest_f$. We observe a FIR-FP ratio above 1 for the low-competition and high-competition dummies. We observe values below 0 for the COE for the same dummies. Extreme market conditions—either too little or too much competition—are therefore commonly misinterpreted by the model as indicators of success.

Both highly centralized and highly decentralized markets frequently misinterpret the

likelihood of success. However, centralized markets may overestimate success due to robust subsidies and support systems that inflate expectations, while decentralized, high-competition markets likely err due to a saturated market with overly optimistic distribution projections without sufficient backing from production or market reception. Conversely, medium competition levels present more accurate predictive outcomes. This is further corroborated by analyzing the interdependence between the competitive environment and the *Distribution* variable (see Figure A.7 in the online Appendix), for which we observe that more distribution in both low and high competition environments does not appear to increase the likelihood of success.

5 Conclusion

In this study, we have examined the impact of critical reviews on predicting commercial success. In contrast with the industry press and audience ratings, *Cineuropa* reviews display a positive bias. This bias can be seen from the false positive predictions in all the models we tested. It generates misleading information that does not correspond to the dynamics of the European film market. These results are robust with regards to the alternative classification method proposed by EAO 2023 Report. The current surge in online services and streaming platforms means that commercial success cannot be measured by attendance figures alone. Future research could improve on the method we use to classify commercial success by taking online viewers into account.

There are several ways to interpret the above findings. On the one hand, critics' tastes may not correspond to those of viewers, who prefer different films. On the other, critics may choose or be encouraged to exclusively review films they like, in hope of better results, be they commercially or artistically motivated. This promotional logic is concerning in several ways. First, it clashes with the 'consumer guiding' role of critics, which is to reward certain films

and reject others, so that viewers can make informed decisions. Second, positive feedback does not translate into increased success, as the number of winners is limited. Assuming that participation in top-tier film festivals is also highly selective, the same winner-takes-all dynamics seen in commercial success also apply to artistic merit.

In another interpretation, *Cineuropa*'s publication practices closely align with the policies of the institutions that fund it. Funding institutions, whether at the regional, national or supranational level, determine the type of films that get made. When supposedly independent market players, such as film critics, receive funding from the same source, their impartiality as experts gets compromised. They may produce wrong incentives that encourage the production of films that serve the priorities of the funding institutions instead of what artists or audiences want. Although we only provide empirical evidence on the 'lemons' aspect of the European film market, Akerlof's model implies that in such markets, actors that propose high-quality products—'better' films in this case—are gradually crowded out. Centralized decision-making might thus undermine the freedom of expression of film critics and filmmakers alike.

For the sake of clarity, we do not advocate for replacing puplic support for filmmakers by market forces. However, we believe that a debate on the future of European cinema is overdue. Film journalism must retain its critical role and not be replaced by promotional or advertising strategies. Similarly, film initiatives and artistic creativity in Europe might benefit from looser institutional control, which currently extends to all aspects of filmmaking, be it the writing, production, distribution or reception of films. If "diversity" is to be an enacted goal of the EU's cultural policy, the current institutional framework works against it and ought to be addressed.

References

Abidi, S. M. R., Y. Xu, J. Ni, X. Wang, and W. Zhang (2020). Popularity Prediction of Movies: From Statistical Modeling to Machine Learning Techniques. *Multimedia Tools and Applications 77*, 35583–35617.

Adler, A. I. and A. Painsky (2022). Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy 24* (5:687).

Ahmed, U., H. Waqas, and M. T. Afzal1 (2020). Pre-production Box-Office Success Quotient Forecasting. *Soft Computing 24*, 6635–6653.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics 84* (3), 488–500.

Basuroy, S., S. Chatterjee, and S. A. Ravid (2003, October). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing 67*, 103–117.

Basuroy, S., S. A. Ravid, R. T. Gretz, and B. J. Allen (2020). Is Everybody an Expert? An Investigation into the Impact of Professional versus User Reviews on Movie Revenues. *Journal of Cultural Economics 44*, 57–96.

Batista, G. E. A. P. A., A. L. C. Bazzan, and M. C. Monard (2003). Balancing Training Data for Automated Annotation of Keywords: a Case Study. In S. Lifschitz, N. F. A. Jr., G. J. P. Jr., and R. Linden (Eds.), *WOB*.

Bridges, J. and C. Vásquez (2016). If Nearly All Airbnb Reviews are Positive, Does That Make Them Meaningless? *Current issues in tourism 21* (2), 1–19.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002, June). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Cook, P. J. and R. H. Frank (1995). *Blockbusters: Hit-making, Risk-taking, and the Big Business of Entertainment*. Free Press.

Council of the European Union, E. P. (2024). Directive (EU) 2024/825 Amending Directives 2005/29/EC and 2011/83/EU as Regards Empowering Consumers for the Green Transition through Better Protection Against Unfair Practices and Through Better Information. https://www.europeansources.info/record/?p=525524.

Creative Europe, M. (2016-2024). Selected Creative Europe Projects. `culture.ec.europa.eu/creative-europe/projects`. Accessed: 2024-09-30.

Dai, X., I. Chalkidis, S. Darkner, and D. Elliott (2022). Revisiting Transformer-based Models for Long Document Classification. arXiv preprent.

Debande, O. (2017). Film Finance: The Role of Private Investors in the European Film Market. In P. C. Murschetz, R. Teichmann, and M. Karmasin (Eds.), *Handbook of State Aid for Film*, pp. 51–66. Media Business and Innovation.

Delen, D., R. Sharda, and P. Kumar (2007). Movie forecast Guru: A Web-based DSS for Hollywood managers. *Decision Support Systems 43*, 1151–1170.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Du, J., H. Xu, and X. Huang (2014). Box Office Prediction Based on Microblog. *Expert Systems with Application 41*(4), 1680–1689.

e Souza, T. L. D., M. Nishijima, and R. Pires (2023). Revisiting Predictions of Movie Economic Success: Random Forest Applied to Profits. *Multimedia tools and applications 82*, 38397–38420.

Elberse, A. (2013). *The Winner-Take-All Society: Why the Few at the Top Get So Much More Than the Rest of Us*. New York: Henry Holt and Co.

Eliashberg, J. and S. M. Shugan (1997). Film Critics: Influencers or Predictors? *Journal of Marketing 61*(2), 68–78.

European Commission, C. (2014). European film in the digital era: Bridging cultural diversity and competitiveness. `eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52014DC0272`. Accessed: 2024-09-29.

European Commission, C. (2020). Europe's Media in the Digital Decade: An Action Plan to Support Recovery and Transformation. `eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0784&from=EN`. Accessed: 2024-09-29.

Gemser, G., M. V. Oostrum, and M. A. A. M. Leenders (2007). The Impact of Film Reviews on the Box Office Performance of Art House versus Mainstream Motion Pictures. *Journal of Cultural Economics 31*, 43–63.

Hartmann, J., M. Heitmann, C. Siebert, and C. Schamp (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing 40*(1), 75–87.

Hassija, V., V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. S. Mahmud, and A. Hussain (2023, aug). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*.

He, H. and E. A. Garcia (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering 21*(9), 1263–1284.

Heinkel, R. (1981). Uncertain Product Quality: The Market for Lemons with an Imperfect Testing Technology,. *The Bell Journal of Economics 12*(2), 625–636.

Hur, M., P. Kang, and S. Cho (2016). Box-office Forecasting Based on Sentiments of Movie Reviews and Independent Subspace Method. *Information Sciences 371*, 608–624.

Johnson, J. M. and T. M. Khoshgoftaar (2019). Survey on Deep Learning with Class imbalance. *Journal of big data 6*(1), 1–54.

Kanzel, M. (2018). Fiction film financing in Europe: A Sample Analysis of Films Released in 2016. *European Audiovisual Observatory*.

Kanzel, M. (2019). Fiction Film Financing in Europe: A Sample Analysis of Films Released in 2017. *European Audiovisual Observatory*.

Kanzel, M. (2020). Fiction Film Financing in Europe: A Sample Analysis of Films Released in 2018. *European Audiovisual Observatory*.

Kanzel, M. (2022). Fiction Film Financing in Europe: A Sample Analysis of Films Released in 2020. *European Audiovisual Observatory*.

Kanzel, M. (2023). An Analysis of European Box Office Structure 2010-2022. *European Audiovisual Observatory*.

Kanzel, M. (2024). Fiction Film Financing in Europe: A Sample Analysis of Films Released in 2021. *European Audiovisual Observatory*.

Kanzel, M. and P. Simone (2022). World Film Market Trends. *Marché du Film*.

Kim, S. H., N. Park, and S. H. Park (2013). Exploring the Effects of Online Word of Mouth and Expert Reviews on Theatrical Movies' Box Office Success. *Journal of Media Economics 26*(2), 98–114.

Kim, T., J. Hong, and P. Kang (2015). Box Office Forecasting Using Machine Learning Algorithms Based on SNS Data. *International Journal of Forecasting 31*(2), 364–390.

Lash, M. T. and K. Zhao (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. *Journal of Management Information Systems 33*, 874–903.

Lee, K., J. Park, I. Kim, and Y. Choi (2018). Predicting Movie Success with Machine Learning Techniques: Ways to Improve Accuracy. *Information Systems Frontiers 20*, 577–588.

Light, D. W. and J. R. Lexchin (2021). Pharmaceuticals as a Market for "Lemons": Theory and Practice. *Social Science Medicine 268*, 113368.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint.

Lundberg, S. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011, June). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Meiseberg, B. and T. Ehrmann (2013). Diversity in teams and the success of cultural products. *Journal of Cultural Economics 37*, 61–86.

Pang, B. and L. Lee (2005, June). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales]. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 115–124.

Pang, B., L. Lee, and S. Vaithyanathan (2002, July). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 79–86.

Ramon, Y., D. Martens, F. Provost, and T. Evgeniou (2020). A Comparison of Instance-level Counterfactual Explanation Algorithms for Behavioral and Textual Data: SEDC, LIME-C and SHAP-C. *Expert Systems with Application Advances in Data Analysis and Classifation*(14), 801–819.

Ravid, S. A., J. K. Wald, and S. Basuroy (2006). Distributors and Film Critics: Does it Take two to Tango? *Journal of Cultural Economics 30*(3), 201–218.

Reinstein, D. A. and C. M. Snyder (2005). The Influence of Expert Reviews on Consumer Demand for Experience Goods. *Journal of Industrial Economics 53*(1), 27–51.

Research and Markets (2024). Europe Movie Market, Size, Forecast 2024-2030, Industry Trends, Share, Growth, Insight, Impact of Inflation, Company Analysis. *Research and Markets*. Accessed: 2024-09-29.

Shapley, L. S. (1951). Notes on the N-Person Game—ii: The Value of an N-Person Game. *The Rand Corporation*.

Sharda, R. and D. Delen (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications 30*, 243–254.

Skowronski, J. J. and D. E. Carlston (1989). "Negativity" and Extremity Biases in Impression Formation: A Review of Explanations. *Psychological Bulletin 105*, 17–22.

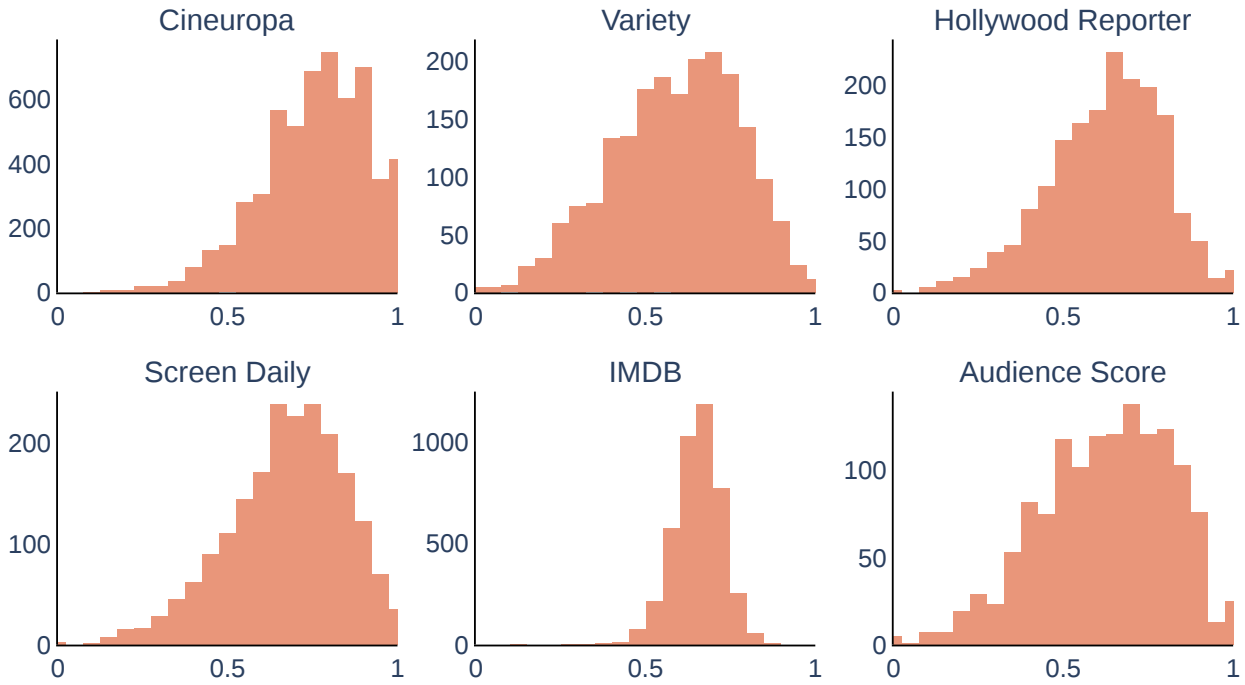
Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics SMC-6*(11), 769–772.

Wallentin, E. (2016). Demand for Cinema and Diverging Tastes of Critics and Audiences. *Journal of Retailing and Consumer Services 33*, 72–81.

Wiggins, Rosalind Z.and Bennett, R. L. and A. Metrick (2019). The Lehman Brothers Bankruptcy D: The Role of Ernst Young. *ournal of Financial Crises 1*, 100–123.

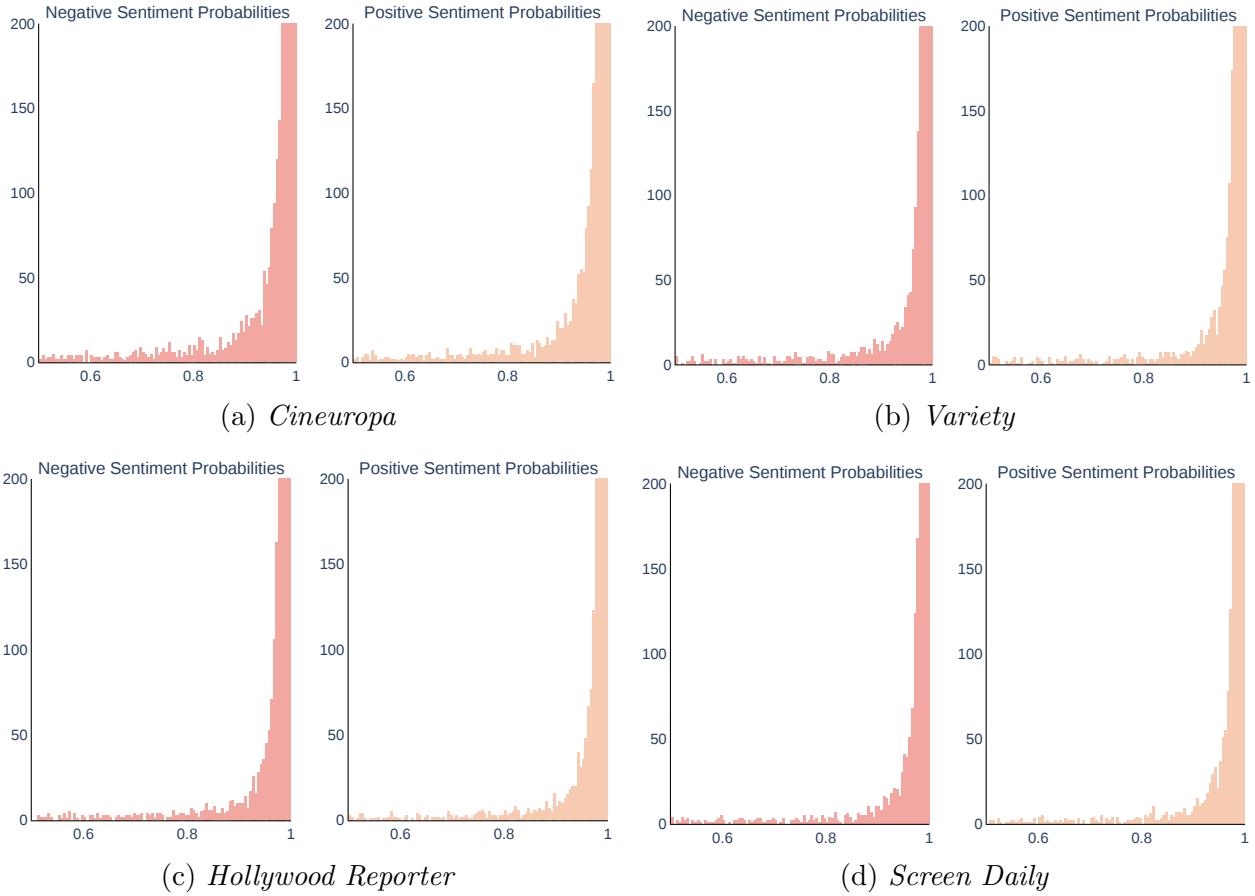
Appendix A: Additional Figures and Tables

Figure A.1: Sentiment Scores of Media Outlets



Note: Distribution of sentiment scores of a data set of film reviews. Sentiment scores are obtained using SiEBERT (Hartmann et al. 2023), where 0 is negative and 1 is positive. All datasets skew positively with averages above 0.5, but this doesn't mean the reviews are necessarily positive. Few negative sentences in the text may have a strong impact on the way a reader perceives the review.

Figure A.2: Probability Distributions of Classification Task

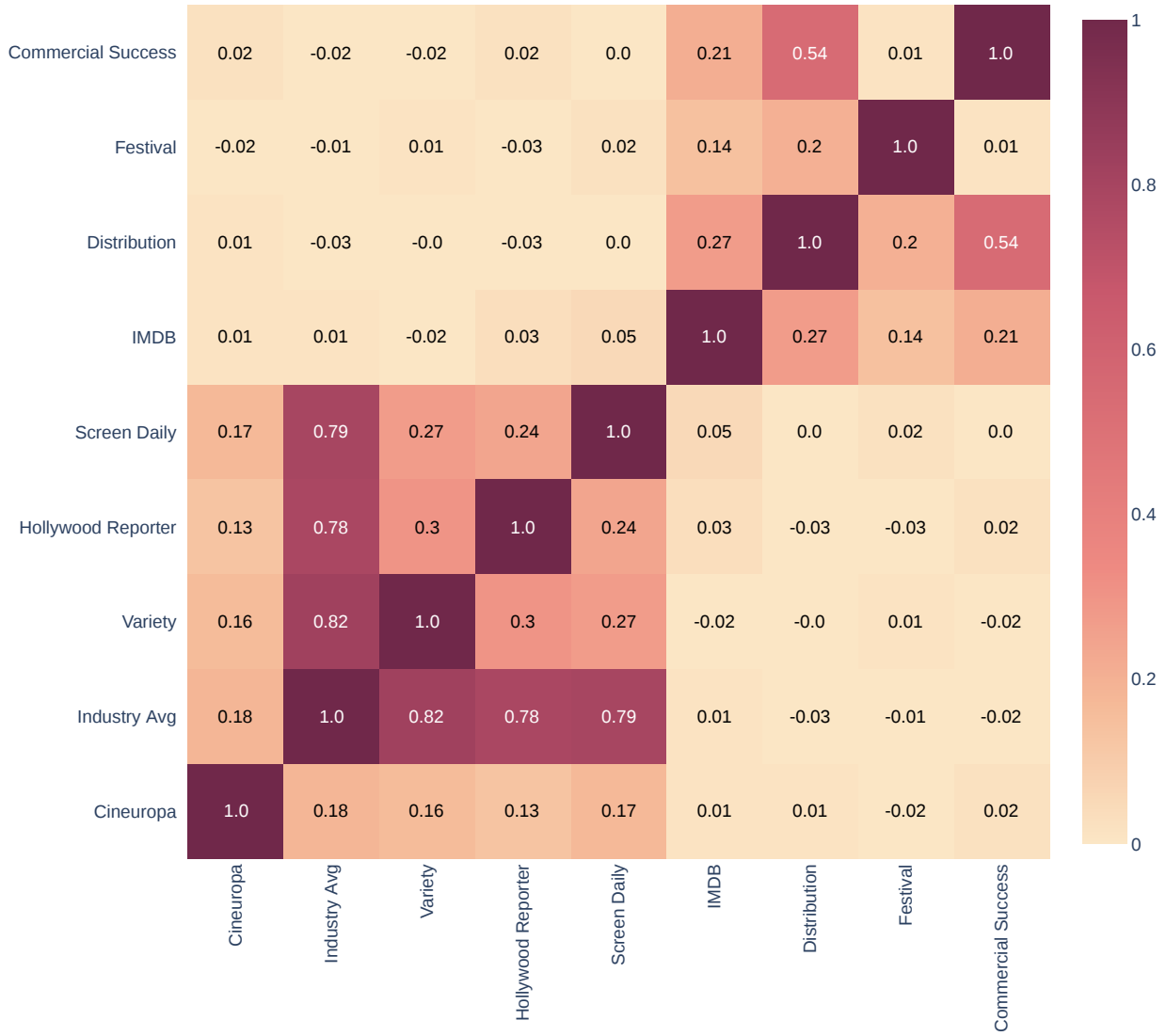


(a) *Cineuropa* (b) *Variety*

(c) *Hollywood Reporter* (d) *Screen Daily*

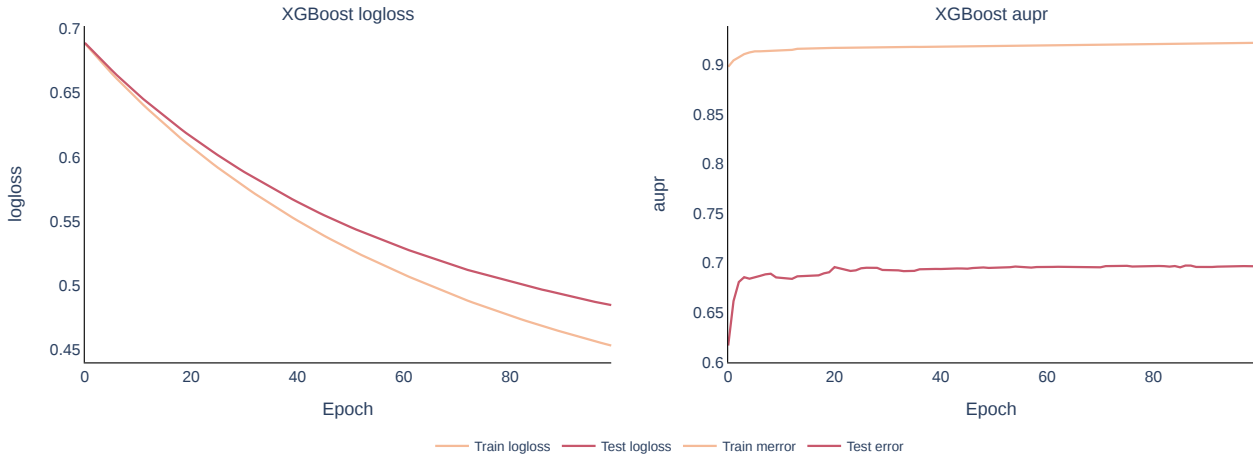
Note: Probability distributions for classifying *Cineuropa* and corresponding industry reviews in *Variety*, *Hollywood Reporter*, and *Screen Daily* using SiEBERT-large (Hartmann et al. 2023). The x-axis has been limited to 200 and the y-axis to 0.5, to make lower probability scores visible. The overall shape of the probability distribution shows that the model performs well, with a slightly better performance for positive than negative classes.

Figure A.3: Correlation Matrix



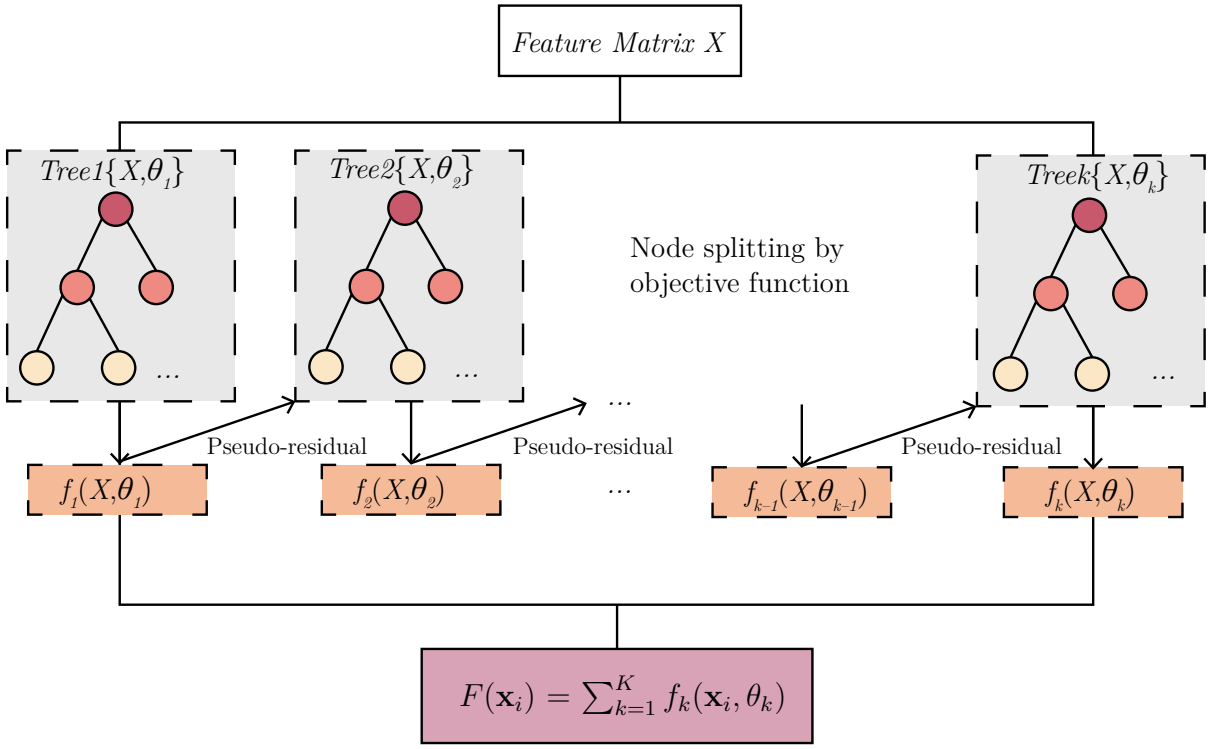
Note: This correlation matrix visualizes the relationships between various film review sources, distribution metrics, and festival participation in predicting the commercial success of European films. The rows/columns include sentiment scores from *Cineuropa*, and industry sentiment scores from *Variety*, *Hollywood Reporter*, and *Screen Daily*, alongside the audience scores from IMDb, a distribution proxy score (Distribution), festival participation (Festival) and commercial success, our target variable. There is a general trend of low correlations between critique scores and commercial success. The strongest predictors of commercial success appear to be the distribution metrics and audience scores (IMDb).

Figure A.4: Learning Curves for Baseline XGBoost



Note: *Figure left* (XGBoost logloss): The curve displays a steady decline, indicating the model improves its training predictions. The test logloss decreases more slowly and remains close to the training curve, suggesting good generalization and minimal overfitting. *Figure right* (XGBoost aupcs): The Train Area Under the Precision-Recall Curve (AUPR) increases steadily, indicating that the model is improving its ability to distinguish between classes on the training data. The test AUPR also increases, though it does so more slowly and eventually stabilizes at around 0.7, indicating no improvement in predicting the positive class during training.

Figure A.5: Sequential Process of Gradient Boosting



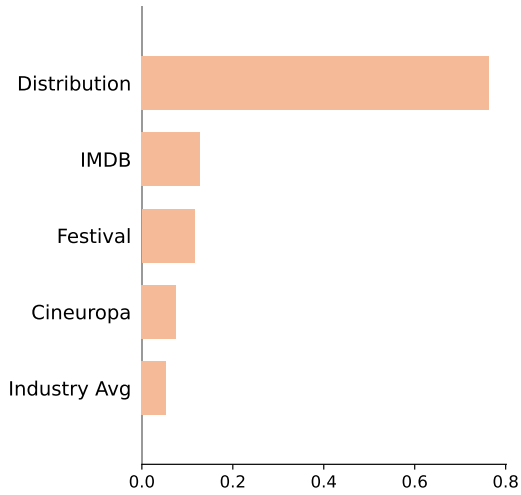
Note: the sequential process of Gradient Boosting as implemented in XGBoost uses decision trees as weak learners. The feature matrix X represents the entire dataset, where each tree $f_k(X, \theta_k)$ is trained on the residuals (errors) from the previous trees to progressively improve the model's accuracy. The sum of the outputs from all trees forms the final model, $F(X) = \sum_{k=1}^{K} f_k(X, \theta_k)$, which is then transformed into predicted probabilities using the *sigmoid* function. This iterative process is effective in handling imbalanced data by prioritizing the reduction of errors associated with minority classes.

Table A.1: Classification EAO 2023 Report Method

Class	Precision	Recall	F1-Score	Support
0	0.97	0.81	0.88	960
1	0.23	0.67	0.34	81
Accuracy			0.80	1041
Macro Avg	0.60	0.74	0.61	1041
Weighted Avg	0.91	0.80	0.84	1041

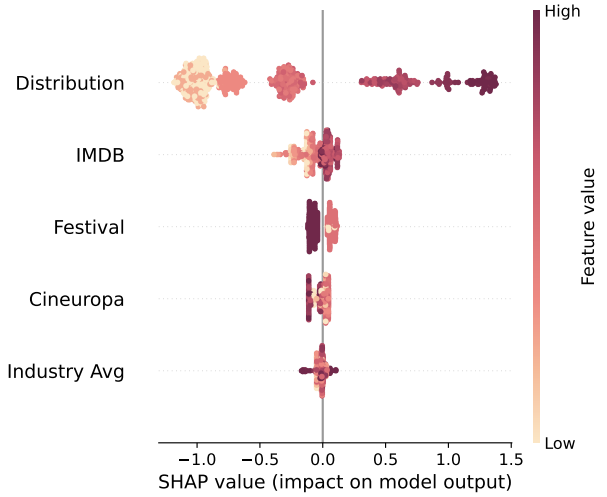
Note: Performance metrics of our baseline model for predicting commercial success of European films across four classes as calculated according to Kanzel (2023) from audience scores, distribution and festival attendance. The metrics include precision, recall, F1-score, and support numbers for each class. Class '0', representing the least successful films, displays the strongest results, while class '1' shows a markedly weaker performance.

Figure A.6: Average Impact on Model Output Magnitude



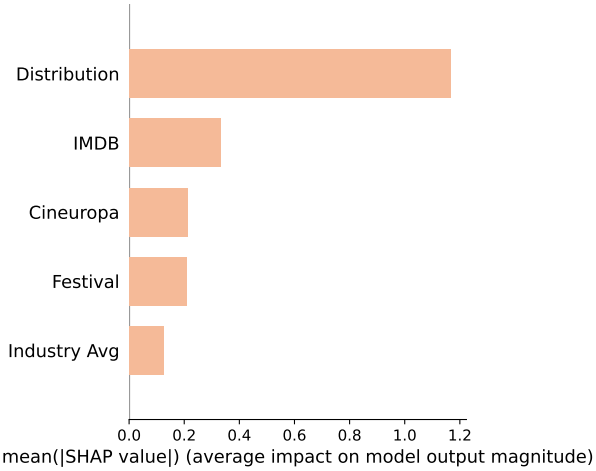
(a) Mean Absolute Feature Impact

Note: Mean absolute SHAP values for the feature variables predicting the commercial success of European films. The model used for these predictions is based on the classification method detailed in Section 3.



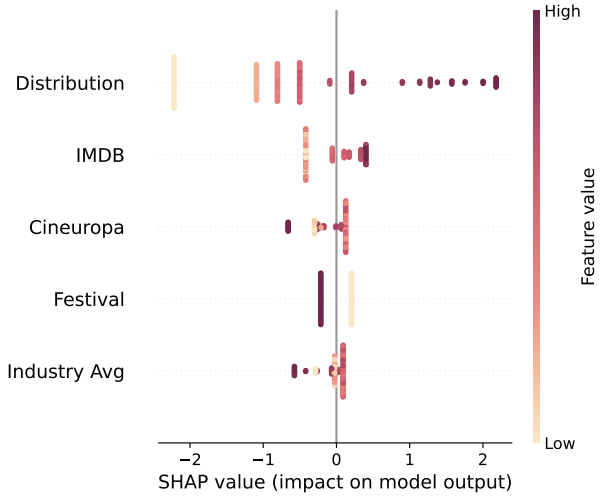
(b) Shap Values Feature Impact

Note: The mean absolute SHAP values for different feature variables used to predict the commercial success of European films. The *Distribution*, and *IMDb* variables show that higher values lead to higher prediction outcomes.



(c) Mean Absolute Feature Impact (EAO 2023 Report)

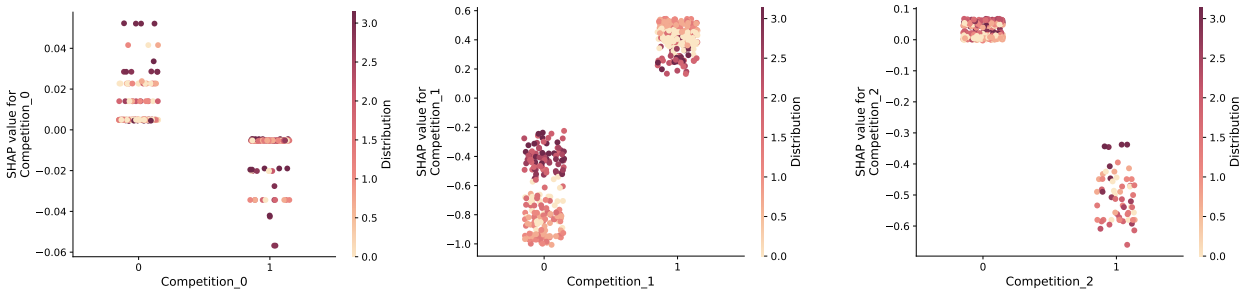
Note: Same as (a) using a model with a different success classification method from the 2023 Report of the European Audiovisual Observatory (Kanzel 2023).



(d) Shap Values (EAO 2023 Report)

Note: Same as (b) using a model with a different success classification method from the 2023 Report of the European Audiovisual Observatory (Kanzel 2023).

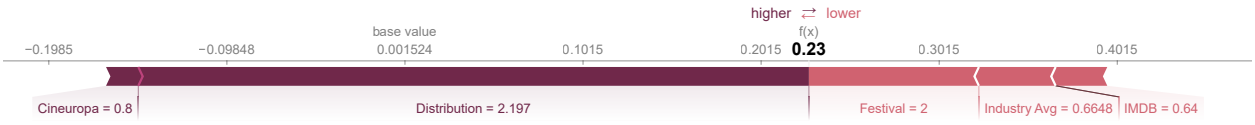
Figure A.7: Dependence Plots Competition Dummies and Distribution



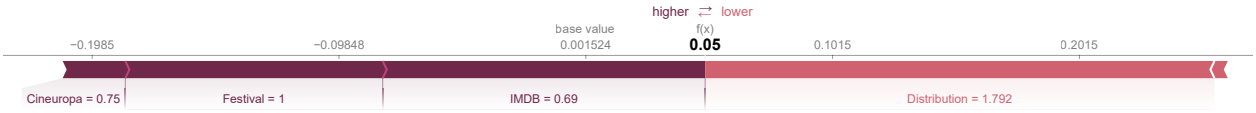
(a) *Low Competition* (b) *Medium Competition* (c) *High Competition*

Note: SHAP dependence plots for competition environment and distribution. Figure (a) shows the values for low *Competition* ($x = 1$). The SHAP values have a slight negative effect, in particular when *Distribution* is high, i.e. the combination of low *Competition* and high *Distribution* decreases the likelihood of predicting a film as successful. Figure (b) shows the values for medium *Competition* ($x = 1$) showing a positive influence on the likelihood of success. Figure (c) shows the values for high *Competition* ($x = 1$) with negative SHAP values for increasing *Distribution*.

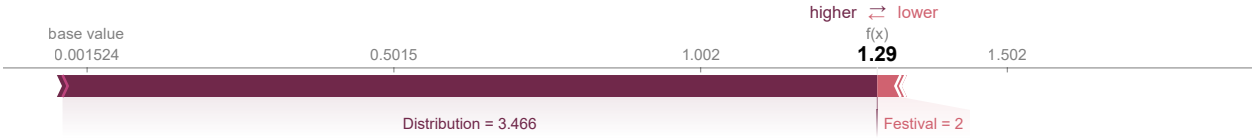
Figure A.8: SHAP Single Predictions



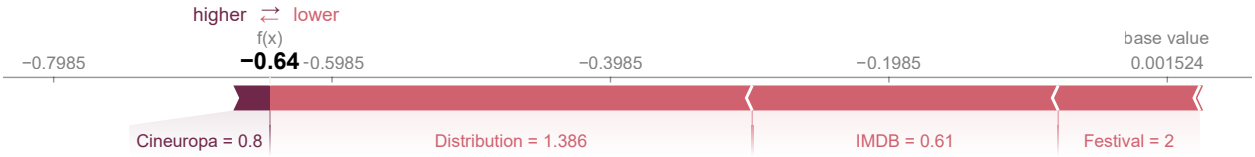
(a) *False Positive*



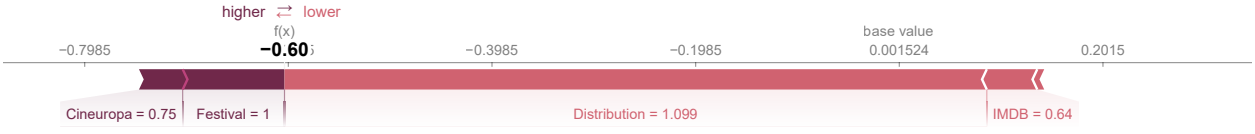
(b) *False Positive*



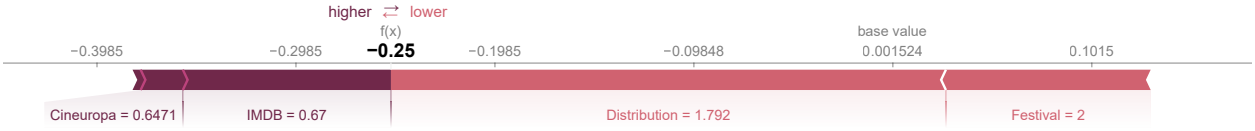
(c) *True Positive*



(d) *True Negative*



(e) *True Negative*



(f) *False Negative*

Note: SHAP force plots for six different cases classified by the baseline model. The cases include (a) and (b) False Positives, (c) True Positive, (d) and (e) True Negatives, and (f) False Negative. Each plot visualizes the contribution of each feature to the model's prediction for a specific instance. Darker colors push the prediction higher (towards success), lighter color push the predictions lower (towards failure). The length of the corresponding bar reflects the magnitude of each contribution.

Appendix B: Sentiment Analysis

B.1 Hyperparameters and Hardware

We use use SiEBERT-large (Hartmann et al. 2023) in the Google Colab environment. We classified our data sets on the sentence-level. For *Cineuropa*, we labeled 101,647, for *Variety* 53,741, for *Hollywood Reporter* 51,226 and for *Screen Daily* 52,727 sentences. On a Tesla T4 GPU with 16GB high-bandwidth memory (GDDR6), this process took about 20-30 minutes for each data set, i.e. around 2 hours for the entire 259,341 sentences. The SiEBERT-large model was fine-tuned using a learning rate of 2e-5, over 3 epochs, with a warm-up phase of 500 steps, a weight decay set to 0.01, and a batch size of 8, which are the default setups of the model.

B.2 Sentiment Classification

Of the 259,341 labeled sentences of all data sets, we only select sentences over 10 words. This is to make sure that sentences with too few words are not taken into account. Next we define a probability threshold of 0.98 for positive and 0.8 for negative sentiment. The reason for the higher threshold for positive sentiments is that the model appears to be better at labeling positive sentiments than negative ones (see Figure A.2). Furthermore, manual verification has shown that descriptive sentences are often labeled as positive. This should not influence the results, however, since it would only raise the overall level of positivity when all negative sentences are still captured. Table B.1 provides an overview of the mean probabilities for each class and data set.

Table B.1: Probability Distribution of Classification Task

Medium	N. (obs)		Class. Probabilities	
	Negative	Positive	Negative	Positive
Cineuropa	22,669	69,452	0.9883	0.9949
Variety	18,506	27,844	0.9908	0.9934
Hollywood Reporter	16,277	28,022	0.9903	0.9936
Screen Daily	13,198	28,129	0.9903	0.9943

Note: Overview probability tensors for classifying *Cineuropa* and corresponding industry reviews in *Variety*, *Hollywood Reporter*, and *Screen Daily* using SiEBERT-large (Hartmann et al. 2023). The Class. Probabilities column reports the mean probability of sentences in each class.

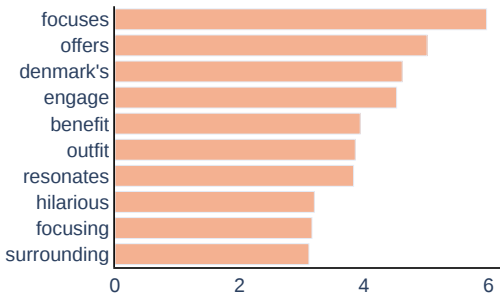
B.3 Explaining SiEBERT Sentiment Analysis with Captum's Layer Integrated Gradients (LIG)

Recent studies on deep-learning and machine learning models seek to better understand how different model inputs determine outcomes (Binder et al. 2016, Shrikumar et al. 2017, Nori et al. 2019, Klaise et al. 2021). To understand SiEBERT's decision-making process, we need to examine how each word influences its decision about the sentiment of the sentences we want it to classify. We use the Layer Integrated Gradients (LIG) module included in the Captum package (Kokhlikyan et al. 2020). The LIG method (Sundararajan et al. 2017) uses attribution scores to explain the relative importance of each input feature in the model's decision. It has two steps. First, a baseline input is determined for which the model's prediction is near zero. In Captum's LIG implementation, the default value is a baseline of zeros (all-zero vector). This is a common convention because the all-zero vector represents a 'neutral' input in many text-based models. Second, the attribution score for an actual input is computed by integrating the gradients of the model's output along a straight-line path from the baseline to the actual input, approximated using a summation over discrete steps. Sundararajan et al. (2017) find that somewhere between 20 and 300 steps are enough to approximate the score. We use 50 steps which is the default value in Captum for this parameter.

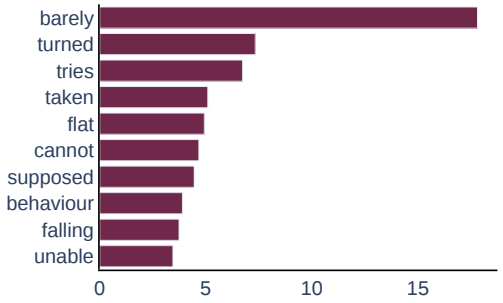
Figure B.1 shows the ten words with the highest positive and negative attribution scores for each outlet. The majority of words are qualitative adjectives (e.g. 'richly', 'terrible') or verbs (e.g. 'offers', 'losing') often used in critical texts, with a few exceptions (e.g. 'Denmark', 'focus', 'taken') drawn from the more descriptive parts of the reviews, such as synopsis, plots or production details. These results suggest that the SiEBERT model is able to interpret the text and identify which parts are likely to carry the most weight in the context of a critical film review.

Figure B.2 shows the Kernal Density Estimation (KDE) of the attribution scores calculated for each outlet. To compare the attribution scores for each outlet, we mean-centered the values for each data set (outlet). We observe that *Cineuropa* has the narrowest and *Variety* the broadest distribution concentration in the values of the attribution scores. Compared with the industry reviews, most tokens used by *Cineuropa* do not express a clear sentiment and thus receive low attribution scores in SiEBERT's assessment.

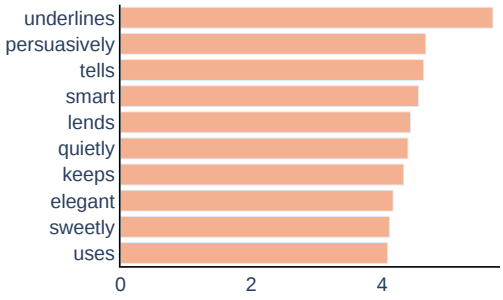
Figure B.1: Top 10 Token Attributions



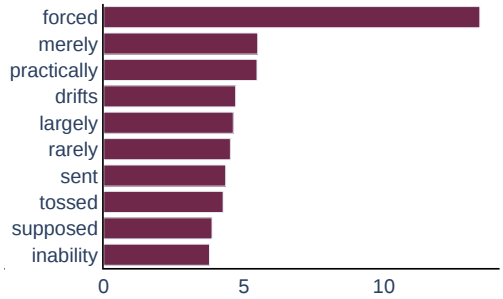
(a) Positive *Cineuropa*



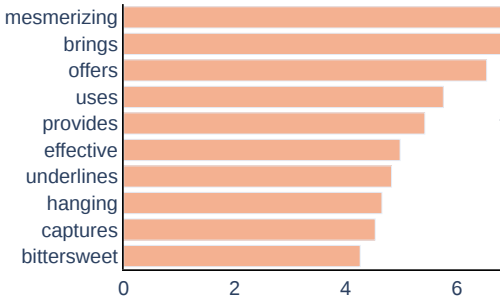
(b) Negative *Cineuropa*



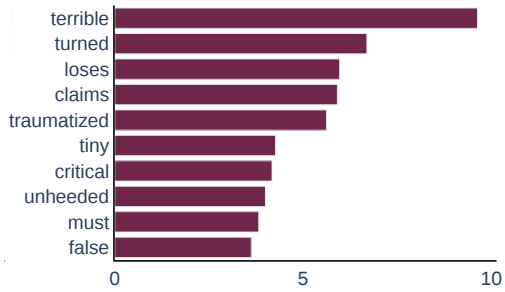
(c) Positive *Variety*



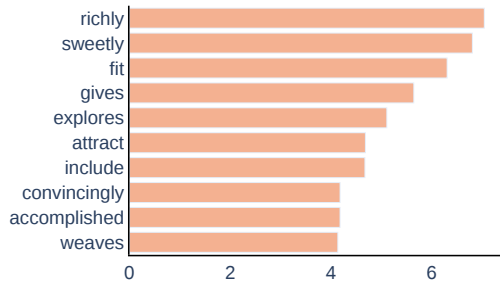
(d) Negative *Variety*



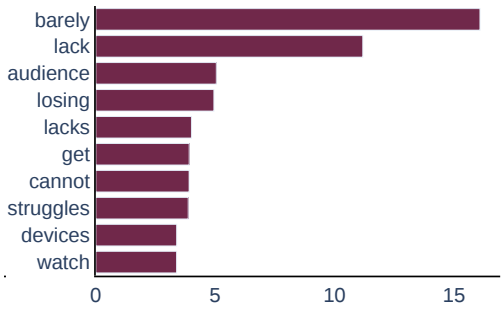
(e) Positive *Hollywood Reporter*



(f) Negative H*ollywood Reporter*



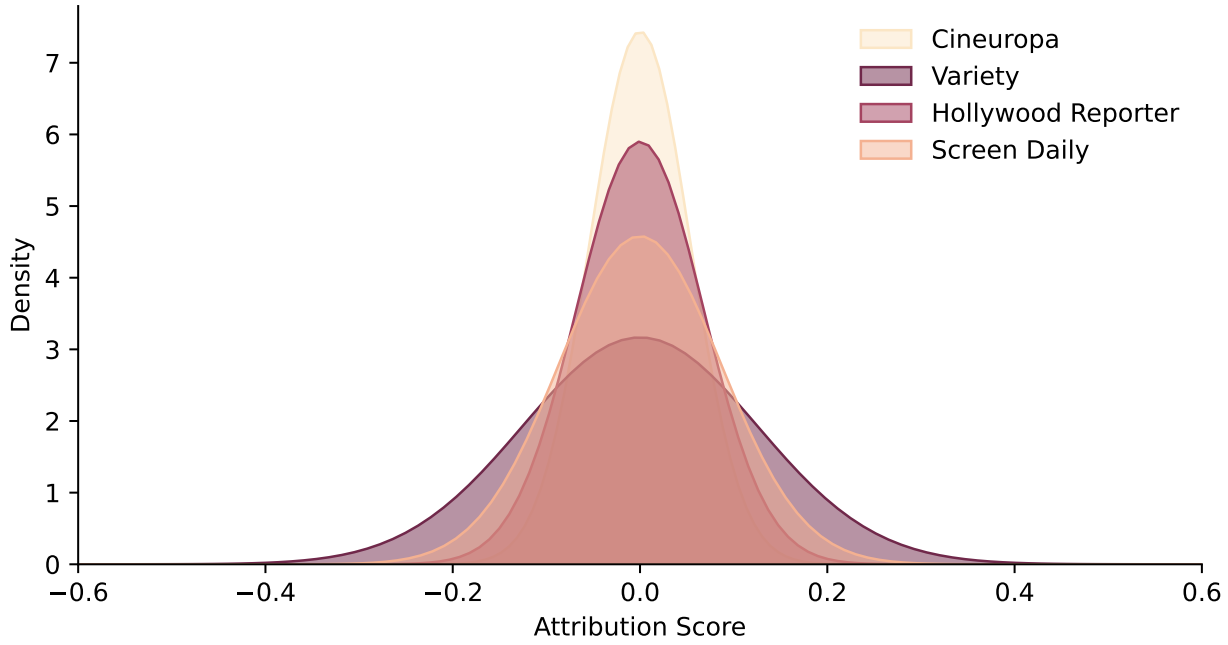
(g) Positive *Screen Daily*



(h) Negative *Screen Daily*

Note: The ten words with the highest positive and negative attribution scores for each outlet. The higher a word's attribution score, the greater its impact on the model's decision. Most words are qualitative terms typical of reviews expressing positive or negative opinions.

Figure B.2: Kernel Density Estimate (KDE) of Attribution Scores



Note: The KDE distribution of the mean-centered attribution score values for each outlet. The curve of *Cineuropa* is centered around 0. This means that most of the words used in *Cineuropa* reviews have a limited impact on SiEBERT's sentiment assessment. *Variety* has the broadest distribution of attribution scores, i.e. a larger proportion of tokens have significant positive or negative influence.

Appendix C: Binary Classification Problem with Imbalanced Data

C.1 SMOTETomek

Class imbalances can skew model performance toward the majority class. In handling this problem, oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) can unintentionally magnify noise and outliers, while undersampling methods may risk discarding information. A hybrid approach is provided by Batista et al. (2003), who found that mixed sampling techniques outperform singular methods by maintaining crucial data features while also improving representational fairness across classes, particularly in severely imbalanced scenarios. Here, we briefly explain the rationale behind SMOTETomek, which combines SMOTE and Tomek Links.

SMOTE is an oversampling method developed by Chawla et al. (2002). Unlike basic random oversampling, which duplicates examples from the minority class, SMOTE creates new examples by interpolating between each minority class data point and its nearest neighbors. This process involves selecting a random minority class sample, calculating the Euclidean distance to its nearest neighbors, and generating synthetic samples by interpolating between these points. The method adds new examples to the dataset that are spatially closer to the existing data points of the minority class.

Tomek Links (Tomek 1976) are a modification of the Condensed Nearest Neighbors (CNN) undersampling technique. It focuses on pairs of close instances from different classes, identifying such pairs where each instance is the nearest neighbor of the other. If such a pair (representing both majority and minority class) is identified and there are no closer instances of the same class, it is defined as a Tomek Link and the majority class instance can be removed. This method effectively cleans the data by removing overlapping points between classes, thus sharpening the decision boundary. Formally, let $d(x_i, x_j)$ denote the Euclidean

distance between x_i and x_j, where x_i is a minority class sample and x_j is a majority class sample. A pair (x_i, x_j) is a Tomek Link if no sample x_k satisfies: $d(x_i, x_k) < d(x_i, x_j)$ or $d(x_j, x_k) < d(x_i, x_j)$.

C.2 eXtreme Gradient Boosting (XGBoost)

Our baseline model is based on the XGBoost algorithm (Chen and Guestrin 2016). In this Appendix, we explain the general idea behind the algorithm and the hyper-parameter settings used during the training process. The gradient boosting algorithm was originally developed by Jerome Isaac Friedman (2001). Boosting is an ensemble technique that merges several weak learners to form a robust predictive model. These base models are trained in sequence, each addressing the errors of its predecessors. In gradient boosting, each model predicts the negative gradients of the loss function, effectively making a steepest descent step towards loss minimization with each addition to the ensemble (see Figure A.5). When the weak learners are decision trees, the resulting method is known as gradient-boosted decision trees (GBDT).

C.2.1 Gradient Boosting

In binary classification problems, the *sigmoid* function is used to model the probability that a sample belongs to the positive class

$$p = P(y = 1|\mathbf{x}) = \sigma(F(\mathbf{x})) = \frac{1}{1 + \exp(-F(\mathbf{x}))} \tag{2}$$

where $F(\mathbf{x})$ represents the predicted log odds, i.e., the logarithm of the odds ratio $F(\mathbf{x}) = \log \frac{p}{1-p}$. The initial model $F_0(\mathbf{x})$ is given by the log odds of the positive class in the training set, i.e. the logarithm of the ratio of the number of positive samples to the number of

negative samples. The loss function is the binary log loss

$$\mathbf{L}(y_i, F_{m-1}(\mathbf{x}_i)) = -y_i \log p_i - (1 - y_i) \log(1 - p_i) \tag{3}$$

where $p_i = \sigma(F_{m-1}(\mathbf{x}_i))$ is the predicted probability of the previous ensemble that sample $\mathbf{x}_i$ belongs to the positive class. The gradient of this loss with respect to $F_{m-1}(\mathbf{x}_i)$ can be computed with the chain rule

$$\begin{aligned}
\frac{\partial L(y_i, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)} &= \frac{\partial(-y_i \log p_i - (1 - y_i) \log(1 - p_i))}{\partial p_i} \frac{\partial p_i}{\partial F_{m-1}(\mathbf{x}_i)} \\
&= \left(-\frac{y_i}{p_i} + \frac{1 - y_i}{1 - p_i} \right) \sigma(F_{m-1}(\mathbf{x}_i))(1 - \sigma(F_{m-1}(\mathbf{x}_i))) \\
&= \left(-\frac{y_i}{p_i} + \frac{1 - y_i}{1 - p_i} \right) p_i(1 - p_i) \\
&= -y_i(1 - p_i) + (1 - y_i)p_i \\
&= p_i - y_i
\end{aligned} \tag{4}$$

Since the derivative of the *sigmoid* function is $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, the pseudo-residual at point $\mathbf{x}_i$ is

$$\tilde{y}_i = -\frac{\partial L(y_i, F_{m-1}(\mathbf{x}_i))}{\partial F_{m-1}(\mathbf{x}_i)} = y_i - p_i, \tag{5}$$

i.e. the pseudo-residual is the actual label of $\mathbf{x}_i$) minus the predicted probability of the previous ensemble that this sample belongs to the positive class. The problem with regression trees as base learners is to find the optimal output values y_j for each leaf node

$$\gamma_{jm} = \text{argmin}_\gamma \sum_{x_i \in R_{jm}} \mathbf{L}(y_i, F_{m-1}(\mathbf{x}_i) + \gamma). \tag{6}$$

However, there is famously no closed form solution for finding the optimal output values y_jm for each leaf node in regression trees. One way to approximate the solution is by taking a single iteration of the Newton-Raphson method. In this approach, the goal is to find the minimum of a twice-differentiable function $f : \mathbb{R} \to \mathbb{R}$ by generating a sequence $\{x_k\}$ starting

from an initial guess $x_0 \in \mathbb{R}$. This sequence is constructed using the second-order Taylor approximations of f around the points x_k,

$$f(x_k + t) \approx f(x_k) + f'(x_k)t + \frac{1}{2}f''(x_k)t^2 \tag{7}$$

The next element in the sequence x_{k+1} is chosen as to minimize this quadratic expansion in t. The minimum can be found by setting the derivative of this expansion to 0:

$$\frac{d}{dt}\left(f(x_k) + f'(x_k)t + \frac{1}{2}f''(x_k)t^2\right) = f'(x_k) + f''(x_k)t = 0 \tag{8}$$

Thus, the minimum is achieved solving for $t = -\frac{f'(x_k)}{f''(x_k)}$ and the next iteration in the sequence is given by $x_{k+1} = x_k + t = x_k - \frac{f'(x_k)}{f''(x_k)}$. The second-order Taylor expansion of the loss function around the point $F_{m-1}(\mathbf{x}_i)$ can be expressed similarly,

$$L(y_i, F_{m-1}(x_i) + \gamma) \approx L(y_i, F_{m-1}(x_i)) + \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\gamma + \frac{1}{2}\frac{\partial^2 L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)^2}\gamma^2 \tag{9}$$

Taking the derivative of this expression with respect to γ gives:

$$\frac{\partial L(y_i, F_{m-1}(x_i) + \gamma)}{\partial \gamma} \approx \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} + \frac{\partial^2 L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)^2}\gamma \tag{10}$$

The derivative of the total loss induced by the samples in region R_j is:

$$\frac{\partial}{\partial \gamma}\left[\sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)\right] \approx \sum_{x_i \in R_{jm}} \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} + \sum_{x_i \in R_{jm}} \frac{\partial^2 L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)^2}\gamma \tag{11}$$

Equating this derivative to 0 gives us the optimal output value for region R_j:

$$\gamma_{jm} = -\frac{\sum_{x_i \in R_{jm}} \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}}{\sum_{x_i \in R_{jm}} \frac{\partial^2 L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)^2}} \tag{12}$$

The numerator in equation 12 represents the sum of the derivatives of the loss function at the data points in region R_m (their pseudo-residuals), while the denominator corresponds to the sum of the second derivatives of the loss function at those same points. Taking the first derivative of the log loss of equation 5, the second derivative can be computed by taking the derivative of the first derivative of log loss:

$$\frac{\partial^2 L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)^2} = \frac{\partial[\sigma(F_{m-1}(x_i)) - y_i]}{\partial F_{m-1}(x_i)} = \sigma(F_{m-1}(x_i))(1-\sigma(F_{m-1}(x_i))) = p_i(1-p_i) \quad (13)$$

Therefore, the optimal output for the lead node R_jm can be expressed as:

$$\gamma_{jm} = -\frac{\sum_{x_i \in R_{jm}} (p_i - y_i)}{\sum_{x_i \in R_{jm}} p_i(1-p_i)} = \frac{\sum_{x_i \in R_{jm}} (y_i - p_i)}{\sum_{x_i \in R_{jm}} p_i(1-p_i)} \quad (14)$$

where $\sum_{x_i \in R_{jm}} (y_i - p_i)$ is the sum of the pseudo-residuals for all samples in region R_jm and $\sum_{x_i \in R_{jm}} p_i(1-p_i)$ is the sum of variances of the Bernoulli distributions for the predicted probabilities in the same region. The final expression shows that is a weighted average of the residuals, where the weights are inversely proportional to the variance of the predicted probability, ensuring that regions with higher confidence (lower variance) contribute less to the update.

C.2.2 Hyperparameter Tuning

We use GridSearchCV to optimize our the hyperparameters of our model. GridSearchCV iterates through multiple combinations of parameter tunes and determines which set of parameters performs best for our predictive model. The grid search was set up to evaluate model performance based on balanced accuracy, using a 5-fold cross-validation. Balanced accuracy adjusts for the fact that traditional accuracy measures can be misleading when dealing with highly imbalanced data. It is defined as $\frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$, where where TP (True Positives) and TN (True Negatives) are the correct predictions for the positive and negative classes, respectively, and FN (False Negatives) and FP (False Positives) are the

incorrect predictions. The outcome of the GridSearchCV provided the best set of parameters as reported in Table C.1. Note that the results are slightly higher than in the final model, which may be due to overfitting in GridSearchCV as there was no early-stopping parameter included in this exercise.

Table C.1: GridSearchCV Hyperparameter Settings

	Baseline			
2-5 Parameter	Baseline	Film Reviews	Winner-Take-All	Competition
Alpha	0.01	0.001	0.1	0.01
Lambda	0.01	0.01	0.1	0.01
Learning Rate	0.01	0.01	0.01	0.1
Max Depth	5	3	3	3
Subsample	0.5	0.5	0.5	0.7
Best Accuracy	0.78	0.78	0.86	0.85
	EAO 2023 Report			
Alpha	0.01	0.01	0.1	0.1
Lambda	0.01	0.001	0.1	0.1
Learning Rate	0.1	0.1	0.1	0.3
Max Depth	1	1	1	1
Subsample	0.5	0.5	0.5	0.5
Best Accuracy	0.77	0.77	0.86	0.84

Note: Optimal hyperparameter values and corresponding Best Score from GridSearchCV. The table lists the best values obtained for the hyperparameters alpha, lambda, learning rate, max depth, and subsample during the tuning process of an XGBoost classifier. The final row shows the highest balanced accuracy score achieved with these parameters. From left to right, each new model includes the features of the previous model.

Appendix D: SHapley Additive exPlanations (SHAP)

D.1 Shapley Additive Explanations

Shapley Additive ExPlanations (SHAP) enables local interpretations of predictions by assessing the importance of each feature for individual predictions. It provides a reliable global understanding of a model making it an effective post-hoc interpretative machine learning method. SHAP is based on Shapley values, a cooperative game theory method developed by Lloyd Shapley (Shapley 1951). It is a fair and axiomatically unique method of attributing rewards from a cooperative game. Where a game is a machine learning model, each feature values are players in a game, and the predicted class membership of the sample is the outcome of the game. Shapley value gives a unique solution to fairly attribute the contributions of each player to the outcome of the game. The Shapley value defines the feature importance of feature value i:

$$\phi_i(\text{val}) = \frac{1}{N!} \sum_{S \subseteq \{x_1,\ldots,x_N\} \setminus \{x_i\}} |S|!(|N| - |S| - 1)![\text{val}(S \cup \{x_i\}) - \text{val}(S)], \tag{15}$$

where S where S is a subset of features in the model, $\text{val}(S)$ corresponds to the model output for S, N is the total number of features, and x is the feature values for the sample to be explained, that is $x \equiv \{x_1, \ldots, x_N\} \in \mathbb{R}^N$. The marginal contribution of each feature x_i is computed as $\text{val}(S \cup x_i) - \text{val}(S)$. The weighting of these marginal contributions depends on the number of ways the subset S could be formed before and after the inclusion of x_i, with weights given as $|S|!$ and $(N - |S| - 1)!$ respectively. The process involves summing these contributions across all possible subsets S, then dividing by $N!$ to compute the average.

The Shapley value is a unique solution because it satisfies the axioms of symmetry (or consistency), dummy (or null effect), and additivity (or local accuracy). The principle of symmetry states that if two feature values, x_z and x_k, contribute equally to the model, they

will receive identical Shapley values.

$$\begin{aligned} \text{val}(S \cup \{x_z\}) - \text{val}(S) = \text{val}(S \cup \{x_k\}) - \text{val}(S) \\ \forall S \subseteq \{x_1, \ldots, x_N\} \setminus \{x_z, x_k\} \\ \implies \phi_z(\text{val}) = \phi_k(\text{val}) \end{aligned} \tag{16}$$

The dummy principle implies that any feature value, like x_z, that does not influence the model's output will have a Shapley value of zero.

$$\text{val}(S \cup \{x_z\}) - \text{val}(S) = 0 \quad \forall S \subseteq \{x_1, \ldots, x_N\} \implies \phi_k(\text{val}) = 0 \tag{17}$$

Finally, local accuracy means that the summation of the Shapley values of all feature values in the model equals the model output. This ensures that the combined effect of all feature values on the model's output is equivalent to the total model output minus the output when no feature values are considered, or, mathematically:

$$\sum_{i \in N} \phi_i(\text{val}) = \text{val}(N) - \text{val}(\{\}) \tag{18}$$

D.2 Tree SHAP

Two main problems arise in the computation of Shapley values, 1) dealing with missing features when computing marginal contributions and 2) managing the computational and algorithmic complexity that can become exponential. Firstly, when calculating Shapley values, it is crucial to define how missing features are treated during the computation of marginal contributions. In game theory, all players (features) are typically present, but this is not always the case in machine learning. Kernel SHAP addresses this by simulating missing features through random sampling, substituting missing values with fixed values from other samples. While this approach allows Kernel SHAP to be adaptable across various machine

learning models, it introduces a variance issue due to the randomness of the sampling method.

Secondly, computing Shapley values can be computationally intensive, especially as the number of features grows. Tree SHAP tackles this problem by specifically adapting to the structure of decision trees. It calculates exact Shapley values by excluding the paths in the decision tree that involve the missing features, thus eliminating the variance caused by sampling. Furthermore, Tree SHAP enhances efficiency by tracking tree traversals, which prevents unnecessary recalculations and reduces the computational complexity from an exponential time complexity of $O(TL^2N)$ (where T is the number of trees, L is the maximum number of leaves per tree, and N is the number of features) to a polynomial time complexity of $O(TLD^2)$, where D represents the maximum depth of any tree. This optimization significantly speeds up the calculation of Shapley values, making it more feasible for use with larger datasets and more complex models.

Appendix E: Commercial Success Classification

E.1 Market and Admissions Data

Table E.1 presents the clusters proposed in the annual reports of the European Audiovisual Observatory (EAO) to divide 34 European countries into three market clusters. Due to the lack of available data, the EAO does not present conclusions for all countries each year. We use the three clusters in our study. Each film is assigned to the Small, Medium or Large market cluster according on its production country. In case of co-productions, only the majority production country is considered (the first country credited). We also test a different method, following the classification method proposed by the EAO in its 2023 report. Instead of three, there are six market clusters and, for each one, admissions thresholds are defined for four commercial success class (see Table E.2). Compared to our classification method, the method proposed in the EAO 2023 Report provides similar results without improving the performance of the XGBoost model.

Table E.1: Market Clusters According to the EAO Annual Reports

Market Clusters	**Admissions**
Small market	$<$ 10 million
Medium market	10 – 50 million
Large market	$>$ 50 million

Market Clusters	**Countries**
Small market	BA BG CY, EE FI HR IS LT LV LU, ME, MK MT SI SK
Medium market	AT BE CH CZ DK GR HU IE NL NO PT RO SE
Large market	DE ES FR GB IT PL

Note: EAO employs since 2015 the same method to divide the European production countries into three clusters for fiction films (Kanzel 2018). The countries represented in the annual reports vary according to the available sample data of each year.

Table E.2: Commercial Success Classification Method in the EAO 2023 Report

Annual Adm.	**Countries**	**Blockbuster Films**	**High-gross. Films**	**Mid-tier Films**	**Low-gross. Films**
> 150 m	FR GB & IE	> 3 m	[1 m–3 m]	[300'–1.5 m]	< 300'
90 – 120 m	DE IT ES	> 1.7 m	[800'–1.8 m]	[250'–800']	< 250'
50 – 70 m	TR, PL	> 950'	[400'–950']	[150'–400']	< 150'
30 – 40 m	NL	> 500'	[200'–500']	[100'–200']	< 100'
8 – 20 m	AT BE CH CZ, DK				
	FI GR HU NO PT RO SE	> 270'	[130'–270']	[35' – 130']	< 35'
2 – 7 m	BG EE HR LT LV SI SK	> 100'	[50'–100']	[20'–50']	< 20'

Note: Commercial success thresholds according to the EAO 2023 report. The report notes that the proposed thresholds should be lower for 2020-2022 to account for the COVID19 pandemic during which audience attendance was low. No further instructions are provided for other years.

It is important to note that the five commercial success classes we propose in Table 2 do not allow us to draw conclusions about the financial success for individual films in terms of profit (i.e. how box office gains compare to individual production budgets). Instead they indicate a film's reach, i.e. how its total number of viewers compares to other films. Manual verification confirms that the proposed thresholds correspond to known information about the commercial success of individual films. For instance, Fernando León de Aranoa's *Loving Pablo* (2017) and Pedro Almodóvar's *Pain and Glory* (2019), both included in the EAO list of "Top 20 European Films by Admissions" in their respective years, received a S_c=*very high*. Similarly, Laurent Cantet's *Arthur Rambo* (2021), featured in Ecran Large's list of "The 10 biggest box office failures in France" for 2021, received a S_c=*micro*. A limit of classification methods without real data on the production budget of films is that they fail to correctly account for outliers. Over-performing low-budget films (whose production budget was significantly lower than the mean) and under-performing high-budget films (whose pro-

duction budget was significantly higher than the mean) will not be correctly classified. For example, Pierre Schoeller's *Un roi, une nation* (2018) receives according to our method an S_c=*medium*, for a moderate number of admissions for the French market (363,510 tickets), even though it was a described as a flop considering its production budget of 17 million. *Magnetic Fields* (2020) by Yorgos Goussis, on the other hand, a Greek film with a low production budget (reportedly a budget of only 6500 euros) that ran for 16 weeks, received an S_c=*micro*, even though its number of admissions (27,307 tickets) means that its revenues surpassed its production budget.

E.2 Sample Representativeness

The LUMIIERE database (lumiere.obs.coe.int) provides data for the paid admissions of films released in 34 European countries. We retrieved the market data for over 30,000 films. In the classification with the XGBoost model we only use the films reviewed by *Cineuropa*. Our final sample for which both the market data and critical reception is known consists of 4181 films (see Table E.3). Figure E.1 presents the ratio of the most successful films (class '1') for the over 300,000 titles retrieved from the LUMIERE database. The curve is similar to the results for our sample reported in Figure 2.

Table E.3: Proportion of our Sample Compared to the annual number of European Films with theatrical release

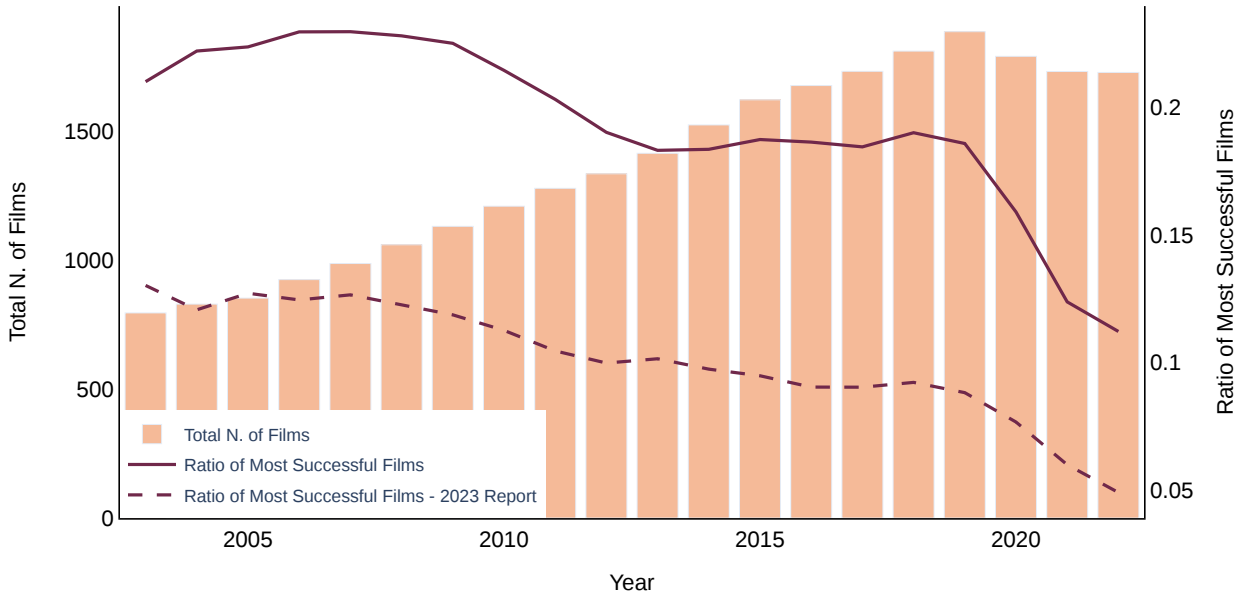
Year	N. (films)	N. (obs)	Proportion
2008	1062	57	0.05
2009	1160	103	0.09
2010	1218	95	0.08
2011	1257	91	0.07
2012	1330	142	0.11
2013	1422	198	0.14
2014	1561	224	0.14
2015	1542	285	0.18
2016	1548	337	0.22
2017	1673	417	0.25
2018	1751	471	0.27
2019	1768	520	0.29
2020	1467	435	0.30
2021	1658	465	0.28
2022	1730	341	0.20
all years	**22,147**	**4181**	**0.19**

Note: Table E.3 shows the representativeness of our sample. The number of films refers to films with a theatrical release each year in at least one of the 34 European countries represented in the LUMIERE database.

E.3 Art vs. Entertainment

Our study focuses on predicting commercial success. While this may be desirable from a financial point of view, the cultural sector may rightly claim that artistic merit is more important than commercial success. Indeed, the European Commission has a stated "joint objective of meeting the challenges of cultural diversity and economic development in the film sector" (European Commission 2014). Cultural diversity, or, artistic success, however, are harder to measure and would require more dynamic factors, for instance an examination of directors frequently winning top prizes at A-list festivals. The main criteria used in our study (e.g. admissions and distribution) are not appropriate to fully account for a film's artistic

Figure E.1: Commercial Success of all films released between 2003 and 2022



Note: Figure E.1 shows the rolling average of the total number of European films released per year and the rolling average of the ratio of the most successful films (class '1'). Class '1' is attributed to films with admissions $A_f > 0.5 \times (B_c/P_c)$, where B_c is the mean production budget and P_c is the average ticket price per year for each market (Class '1' combines the medium, high, and very high success categories S_c, see Table 2). The mean value for the ratio of most successful films released between 2003 and 2022 is 0.17. The dashed line shows the ratio for the most successful films classified according to the EAO 2023 Report (Kanzel, 2023).

potential. The history of cinema counts many artistically innovative and influential films that audiences and critics failed to appreciate in their time. Nevertheless, we believe that our study is also artistically revealing. First and as already mentioned in the introduction, there are niche audiences for art-house films, which means that artistic merit is not antithetical to admission size. While some artistic 'masterpieces' may go unnoticed by audiences, it is important to note that *Cineuropa* mainly reviews 'art-house' films. These are different from 'art' films which address more 'elite' audiences and are, from an economic point of view, not consumer goods but assets. As such, audience size remains important to 'art-house' films (and to EU policymakers). Our metrics—success categories based on admissions—are

therefore able to capture their performance in both artistic and commercial terms. As a reminder, the audience thresholds for our categories are lower than the ones proposed in the EAO 2023 Report and can identify more films as successful.

Recognition by prestigious A-list festivals, such as Cannes, Berlinale or Venice as well as critical acclaim are further markers of a film's cultural significance. For filmmakers, participation in such festivals may be as or even more desirable than commercial success. We use the $Fest_f$ variable to examine artistic success and confirm some of the observations that came from our economic analysis. Festival participation appears to be a key characteristic of our data set. Only 6.6% of films did not attend a festival and most of these are big blockbusters that went to direct release. As our data set includes only films reviewed by *Cineuropa*, this means that film critics rarely report on films that have not been selected in festivals, regardless of their audience figures.

Table E.4: Average Sentiment Scores by Level of Competition

Festival	**Industry Avg**	**Cineuropa**	**N. (obs.)**
Low Competition			
None	0.657931	0.800000	68
B	0.631222	0.785714	578
A	0.643751	0.772727	713
Medium Competition			
None	0.650417	0.800000	183
B	0.656511	0.769231	1109
A	0.645760	0.785714	966
High Competition			
None	0.585507	0.772727	27
B	0.634921	0.766968	232
A	0.664091	0.764706	309

Note: Average sentiment scores for films grouped by competition levels across different festivals. The table compares the sentiment scores assigned by industry averages and *Cineuropa* across three levels of competition. The scores indicate a trend where *Cineuropa* consistently rates films higher in less competitive settings, while industry averages show higher levels of sentiment with higher degrees of competition and festival importance.

Table E.4 reports the festival selection and review scores grouped by their festival participation and level of *Competition* as explained in 4.3. Films in the highest competitive environment show the highest percentage (8.74%) of being selected for A-level festivals and achieving strong critical acclaim (a sentiment scores above 0.8) as opposed to 7.29% for low and 6.52% for medium competition environments, respectively. We observe that in high-competition environments, where productions are often more marginal and receive less funding, films tend to exhibit the highest artistic merit, as can be seen from their selection for prestigious A-list festivals and strong critical acclaim. Conversely, low competition does not appear to lead to higher A-festival selection or better reviews.

If less competitive environments consistently yield lower critical acclaim and fewer selections for top-tier festivals, increasing funding in these contexts may not necessarily lead to greater artistic breakthroughs. On the contrary, an oversupply of less successful films, amplified by misleading signals, may risk the marginalization of the very films that are the most artistically significant.

At the risk of being redundant, we do not want to suggest that limiting public support to create a more competitive environment would necessarily lead to more commercially and artistically successful films. However, the strategy of supporting more and more films in the hope of producing a few winners, as it is the case in low-competition environments, may be hard to justify based on our observations. Engaging in an open dialogue about imbalances in the current support strategies could enhance the diversity of European films.

References Appendix

Batista, G. E. A. P. A., A. L. C. Bazzan, and M. C. Monard (2003). Balancing Training Data for Automated Annotation of Keywords: a Case Study. In S. Lifschitz, N. F. A. Jr., G. J. P. Jr., and R. Linden (Eds.), *WOB*.

Binder, A., G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. *ArXiv abs/1604.00825*.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002, June). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

European Commission, C. (2014). European film in the digital era: Bridging cultural diversity and competitiveness. `eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52014DC0272`. Accessed: 2024-09-29.

Friedman, J. H. (2001). Greedy Function Approximation: a Gradient Boosting Machine. *Annals of statistics*, 1189–1232.

Hartmann, J., M. Heitmann, C. Siebert, and C. Schamp (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing 40*(1), 75–87.

Kanzel, M. (2018). Fiction film financing in Europe: A Sample Analysis of Films Released in 2016. *European Audiovisual Observatory*.

Kanzel, M. (2021). Fiction Film Financing in Europe: A Sample Analysis of Films Released in 2019. *European Audiovisual Observatory*.

Kanzel, M. (2023). An Analysis of European Box Office Structure 2010-2022. *European Audiovisual Observatory*.

Klaise, J., A. V. Looveren, G. Vacanti, and A. Coca (2021). Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research 22*(181), 1–7.

Kokhlikyan, N., V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson (2020). Captum: A unified and generic model interpretability library for pytorch. url/https://arxiv.org/abs/2009.07896.

Nori, H., S. Jenkins, P. Koch, and R. Caruana (2019). Interpretml: A unified framework for machine learning interpretability.

Shapley, L. S. (1951). Notes on the N-Person Game—ii: The Value of an N-Person Game. *The Rand Corporation*.

Shrikumar, A., P. Greenside, and A. Kundaje (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*.

Sundararajan, M., A. Taly, and Q. Yan (2017). Axiomatic attribution for deep networks.

Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics SMC-6*(11), 769–772.