Article

DRECE: Unlocking the Secrets of the Korean Film Market via an AI-Powered Box-Office Classification and Trend Analysis

Subeen Leem [1], Jisong Oh [2], Dayeong So [3] and Jihoon Moon [1,2,3,*]

[1] Department of Medical Science, Soonchunhyang University, Asan 31538, South Korea; qlsl0519@sch.ac.kr
[2] Department of AI and Big Data, Soonchunhyang University, Asan 31538, South Korea; jso2562@sch.ac.kr
[3] Department of ICT Convergence, Soonchunhyang University, Asan 31538, South Korea; sodayeong@sch.ac.kr
* Correspondence: jmoon22@sch.ac.kr

Abstract: The Korean film market has been rapidly growing, and the importance of explainable artificial intelligence (AI) (XAI) in the film industry is also increasing. In this highly competitive market, where producing a movie incurs substantial costs, it is crucial for film industry professionals to make informed decisions. To assist these professionals, we propose DRECE (short for Dimension REduction, Clustering, and Classification for Explainable AI). The DRECE model is an AI-powered box office classification and trend analysis model that provides valuable insights and data-driven decision-making opportunities for the Korean film industry. The DRECE framework starts with transforming multi-dimensional data into two dimensions through dimensionality reduction techniques, grouping similar data points through k-means clustering, and classifying movie clusters through machine-learning models. The XAI techniques used in the model make the decision-making process transparent, providing valuable insights for film industry professionals to improve box office performance and maximize profits. With DRECE, the Korean film market can be understood in new and exciting ways, and decision-makers can make informed decisions to achieve success.

Keywords: box-office; classification; clustering; deep autoencoder; explainable artificial intelligence; machine learning; uniform manifold approximation and projection

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

1. Introduction

Data-driven decision-making (DDDM) has become increasingly important in recent years due to the advancement of technology and the abundance of data available [1–3]. In many industries, people and organizations seek efficient and effective ways to process and analyze large volumes of data to support their decision-making efforts [4–7]. Traditional decision-making methods, which rely on personal knowledge, experience, and wisdom, are limited in their ability to deal effectively with big data and are prone to bias and errors [8, 9]. DDDM involves using models and algorithms to process and analyze different data sources to provide reliable decision support [10, 11]. This approach has been widely applied in various industries, including medical diagnosis [12], financial risk prediction [13, 14], public affairs governance [15], landslide susceptibility prediction [16, 17], travel mode choice [18], and the safe operation of wastewater treatment processes [19, 20], among others [21–23]. DDDM helps reduce the limitations of traditional decision-making methods, resulting in more accurate predictions and better decision support [1, 6, 9, 11]. This can lead to improved efficiency and reduced risks in various industries. Hence, developing new and improved DDDM models and algorithms is a crucial research area and has become a significant trend in decision analysis.

The use of DDDM models in the film industry can also help ensure the industry's steady growth [24, 25]. By providing business decision support and guidance for filmmakers and distributors, DDDM models can help increase the chances of their project's success, drive the industry's continued growth, and ensure its long-term success [26]. For potential investors, using DDDM models can provide valuable information about the potential success of a film before they invest. This can help ensure that their investments are profitable and generate the needed revenue. For directors, DDDM models can provide insights into the success of their films and help them make informed decisions about which production companies to partner with. Hence, the use of DDDM models in the film industry has the potential to play a critical role in reducing the risks associated with the industry and ensuring its sustainable growth. By providing more accurate predictions and better decision-making support, DDDM models can help filmmakers, investors, and directors make informed decisions and improve the efficiency of their operations [27]. The use of DDDM models in the film industry is a clear example of how Industry 4.0 technology can revolutionize traditional industries and drive progress and innovation [26, 27].

Predicting the success of a film at the box office is a complex task due to several factors [28]. The first factor is the short life cycle of films, which are typically screened within three months [29]. This short screening period, combined with the wide range of events that can occur during the screening, makes it difficult to determine which type of marketing will have the most impact on box office success. The second factor is the strong influence of word of mouth (WOM) on the film industry [30]. Unlike other industries, demand for films is heavily impacted by what people say about the film to others. Despite these challenges, researchers have attempted to predict a film's box office success using artificial intelligence (AI), namely data mining, machine learning, and deep learning techniques [31, 32]. These studies have focused on identifying the factors that impact box office success and developing predictive models based on those factors. The studies suggest that big data, such as social media activity [33], Google searches [34], and Wikipedia page activity [35], can instantly predict a film's box office success.

The Korean film market has been gradually growing due to the influence of the Korean Wave [36, 37], and the importance of explainable artificial intelligence (XAI) is also increasing in the AI sector [38, 39]. In the highly competitive film industry, where producing a movie incurs substantial costs, it is crucial for Korean entertainment companies to systematically analyze various factors affecting a movie's box office performance before production [40]. In this context, constructing a box office classification model using XAI can provide valuable insights for film industry experts, who are not necessarily AI specialists, to make informed decisions. As an AI technique, XAI provides accurate predictions and makes it possible to interpret and understand the factors affecting the box office. This can help film industry experts set diverse targets for box office improvement based on movie types, thereby increasing the chances of success. As DDDM models become increasingly common in various industries, incorporating XAI into the decision-making process for the Korean film industry can bring significant benefits and set it apart from its competitors.

This research presents a classification model for box office types of movies to aid decision-makers such as directors, marketers, and production companies in maximizing their profits in the cinema of South Korea. We analyzed the top 300 box office datasets obtained from the Korean Film Council's online database to achieve this. Our data were unclassified, making it challenging to apply supervised learning. As a result, we used the elbow technique to detect clusters with a K-means clustering algorithm, then assigned "1" to each box office in the cluster and "0" to the others to rebuild the dataset. We trained the data through four machine-learning methods and evaluated their performance. Finally, we examined the factors that contributed to each film's box office success and classified the historical box office types. By utilizing XAI, we identified the factors contributing to the success of each box office type, providing valuable insights for decision-makers in the film industry. This study aims to support decision-makers in making informed judgments by systematically classifying box office types and analyzing box office factors.

The main contributions of this study are summarized as follows:

1) We comprehensively understand the Korean film market through our utilization of data collected from the Korean Film Council's online integrated computer network and offer a data visualization approach that incorporates machine learning and data mining techniques to bridge the DDDM of Industry 4.0 and the film industry.

2) By considering various input variables representing movie characteristics, we identify the factors impacting a movie's box office success in South Korea. Our proposed box office classification model is designed to assist film industry professionals in making data-driven decisions to increase the success of future films in the Korean market.
3) By reducing the feature dimension and applying data mining techniques, we effectively cluster movies and analyze box office trends for each cluster. Utilizing XAI, our model interprets the factors affecting box office performance, providing valuable insights for decision-makers in the Korean entertainment industry to improve box office success.

We organize the rest of the paper as follows: In Section 2, we list related studies on the box office. In Section 3, we describe our proposed approach, including the materials and methods used. In Section 4, we present the results of our experiments and engage in discussion. Finally, in Section 5, we conclude our findings and outline potential directions for future research.

2. Related Studies

Previous studies have attempted to predict the success of a movie at the box office by utilizing artificial intelligence techniques like machine learning and deep learning. Table 1 highlights the utilization of AI techniques in these studies. The studies aim to identify the factors affecting box office performance and develop predictive models based on them. The findings suggest that significant data sources are crucial in predicting box office success. The studies have employed different approaches, including machine learning algorithms and factors like reviewing sentiments and social media data to predict box office revenue. In this section, we review the related studies and their findings, highlighting the strengths and limitations of each approach.

Zhang et al. [41] conducted a study to predict the success of a movie at the box office before its theatrical release. They employed a multilayer backpropagation (MLBP) neural network (NN) with multiple inputs and outputs to build the prediction model. The movies were divided into six categories, ranging from "blob" to "bomb," based on their box office revenue. The input variables were selected based on market surveys, and their weight values were determined using statistical methods. The NN was optimized using theoretical guidance and experiments. A classifier with dynamic thresholds was used to standardize the output and improve the robustness of the model. A 6-fold cross-validation experiment was used to measure the prediction model's performance. The results showed that the MLBP model had better prediction accuracy compared to the multilayer perceptron (MLP) method, with 68.1% pinpoint accuracy and 97.1% accuracy within one category. Kim et al. [42] proposed a novel method for predicting film box office earnings using social network service (SNS) data and machine learning algorithms. Three sequential forecasting models were developed to predict non-cumulative and cumulative box office earnings before, one week after, and two weeks after a film's release. SNS mentions, weekly trends, and screening information were used as input variables. A genetic algorithm was used to select significant input variables, and three machine learning-based nonlinear regression algorithms were used to build the forecasting models. The results showed that using SNS data and machine learning algorithms improved the accuracy of all three models. The research process involved selecting films, collecting screening and SNS data, determining the structure of the forecasting models, and selecting input variables. The conclusion was that their new approach, which uses current screening and SNS information, improves the accuracy of forecasting box office earnings.

Hur et al. [43] proposed a new method for forecasting movie box office earnings that takes review sentiment into account and employs non-linear machine learning algorithms. They used viewer sentiments from review texts as input variables in addition to conventional predictors and three machine learning-based algorithms, such as classification and regression tree (CART), artificial neural network (ANN), and support vector regression

(SVR), to capture the non-linear relationship between the box office and predictors. An independent subspace method (ISM) was applied to provide variable importance. The results showed that the proposed methods could make accurate and robust forecasts. A framework for box office forecasting was developed, and experiments were conducted to validate the ISM and verify the predictive performance of the proposed framework. The results showed that the ISM could assess variable importance robustly, and the proposed forecasting framework had good predictive performance. Lee et al. [44] aimed to predict movie box office revenue using ensemble methods. The authors compared the prediction performance of decision trees (DTs), k-nearest neighbors (KNN), and linear regression using ensemble methods, such as random forests (RFs), bagging, and boosting, with a sample of 1439 movies. The results showed that DTs using ensemble methods outperformed KNN and linear regression in predicting box office revenue for the first, second, and third weeks after release. The study also compared the prediction performance between ensemble and non-ensemble methods within each algorithm and found that DTs using ensemble methods provided better application effectiveness than KNN and linear regression analysis. The study was significant as it analyzed Korean movie data, which had rarely been investigated in the movie literature, and provided insights into predicting movie box office revenue using ensemble methods.

Lee and Choeh [45] examined the relationship between movie resource powers and box office revenue and how efficiency moderates the relationship between online word-of-mouth (eWOM) and revenue. Using data envelopment analysis, they found that movie efficiency had negative and positive moderating effects on different eWOM variables and their impact on subsequent box office revenue. Their analysis using DTs, KNN, and linear regression showed that movies with inefficient resources had better prediction performance than movies with efficient resources. The study added to the literature on eWOM by suggesting production efficiency as a moderator between eWOM and the box office. The production efficiency produced by data envelopment analysis (DEA) had yet to be used in previous studies on box office revenue. The authors showed that efficiency could affect the impact of different eWOM variables on the box office. Bogaert et al. [46] aimed to investigate the power of social media data (Facebook and Twitter) in predicting box office sales and determine which platform and data type are the most important. The authors used various prediction algorithms to compare models using movie, Facebook, and Twitter data. The analysis showed that social media data significantly improves the predictive power of traditional box office prediction models, with Facebook data performing better than Twitter data. The sensitivity analysis revealed that the volume and valence-based combination variables of Facebook comments were the most critical variables. The study found that Twitter had less impact on box office sales due to the lower source credibility of Twitter users. The framework employed in the study was based on the cross-industry standard process for data mining (CRISP-DM) methodology. The study results are significant for practitioners, marketers, and academics who want to use social media data for box office sales predictions.

Pan [47] studied the factors affecting the box office revenue of the top 100 films in 2019. The results indicated that score, potential audience, release schedule, and place of origin significantly impacted the box office revenue. The high fit of the model showed that these four independent variables well explained the dependent variable. The model's residuals had weak autocorrelation due to low multicollinearity between the four variables. However, the positive comment rate was found to have no significant impact. The dummy variables revealed that films released on popular schedules had higher box office revenue, and films produced in China outperformed those produced abroad. The study found that film themes significantly impacted the box office revenue, with science-fiction films having the highest average box office. The first-day box office was also found to significantly impact the total box office, reflecting a conformity effect among consumers. Li and Liu [48] researched predicting box-office revenue in the movie industry. They proposed a machine-learning-based method for forecasting box office revenue in the United States (US) and China. The method was tested and compared with eight other methods, and it was found that the support-vector-machine-based method using a gross domestic product (GDP) achieved the best results with a relative root mean squared error of 0.056 in the US and 0.183 in China. The results were validated using data from 2017, 2018, and 2019, and the mean relative absolute percentage errors were found to be 0.044 in the US and 0.066 in China. The study concluded that the proposed method effectively and efficiently predicted nationwide box office revenue. The results provide evidence for the superiority of the support vector machine (SVM)-based method compared to other methods and demonstrate the potential of using economic factors in predicting box office revenue.

Ni et al. [49] aimed to predict the box office revenue of films in China. The authors collected data from ENDATA, which included 5,683 pieces of movie data, and selected the top 2,000 pieces to be used as the prediction dataset. To study the factors influencing the box office, the authors used various types of Chinese microdata, a Baidu search index of movie names, and data on the coronavirus disease 2019 (COVID-19) epidemic. Using a two-layer model architecture, they optimized the stacking algorithm using a machine-learning technique. The base learners were extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), categorical boosting (CatBoost), gradient boosting decision tree (GBDT), RF, and SVR. At the same time, the meta-learner was a multiple linear regression model. The prediction error was 14.49% as measured by the mean absolute percentage error. The results showed that the COVID-19 epidemic at the time of the movie's release had a related impact on the movie's box office. Velingkar et al. [50] studied the film industry, a multi-billion-dollar business that significantly contributes to a country's economy. They focused on the box office revenue of a movie, which is a crucial indicator of its popularity and can be influenced by various factors such as the production company, genre, budget, reviews, and ratings. To help investors make informed decisions, the authors created a machine-learning model that predicts a movie's box office revenue based on information available before its release. The authors used various algorithms, including XGBoost, RF, CatBoost, LightGBM, Ridge, and voting regressors, and considered factors such as the movie's genre, original language, title, popularity rating, release date, budget, cast, crew, and more. The model considers the intended genre and target revenue and uses the RF model to suggest a budget, runtime, star power, and expected popularity that would lead to the desired box office revenue.

While previous studies have also contributed significantly to box office prediction, our research provides a unique perspective on the Korean film industry by classifying box office types and analyzing factors contributing to box office success. This approach differs from previous studies, which mainly focused on predicting box office revenue using machine-learning algorithms. Using data from the Korean Film Council's online database, our study offers a comprehensive understanding of the Korean film market. It employs machine learning and data mining techniques to cluster movies effectively and analyze box office trends for each cluster. The results of our study, including valuable insights into the factors affecting box office performance, can assist decision-makers in the Korean entertainment industry in making data-driven decisions to improve box office success. Additionally, our study uses XAI to interpret these factors, providing a more comprehensive understanding of the Korean film market.

Table 1. Summary of AI-based box office prediction and classification models.

Author(s)	Year	AI Techniques	Input Variables (Features)	Output Variables
Zhang et al. [41]	2008	MLBP neural network	Cinema information, competition, content category, nation, propaganda, showing time, star value	Movie class or performance
Kim et al. [42]	2014	GPR, KNN, MLR, SVR	Number of SNS mentions, screening-related information, weekly trends	Box office earnings
Hur et al. [43]	2016	ANN, CART, ISM, SVR	Movie data, viewer sentiments from review text	Number of audiences
Lee et al. [44]	2020	Bagging, Boosting, DT, KNN, linear regression	Movie-related variables, number of eWOMs	Box office at weeks 1, 2, and 3 after release
Lee and Choeh [45]	2020	Bagging, DEA, DT, KNN, linear regression	Four eWOM (i.e., review depth, review rating, review volume, and the number of positive reviews)	Box office revenue
Bogaert et al. [46]	2021	Bagging, DT, GBM, KNN, linear regression, neural network, RF	Movie data (MOV), MGC and UGC from both Facebook and Twitter	Movie sales data
Pan [47]	2022	ANOVA, regression analysis	Box office, film title, film theme, monthly film box office in 2019, the monthly number of film releases in 2019, number of potential audiences, place of origin, positive comment rate, schedule, score, WOM	Box office revenue
Li and Liu [48]	2022	ARIMA, DNN, linear regression, log-linear regression, ridge regression, RF, SVM	Historical (2002–2010) box office information	China GDP, China NMS, US GDP, US NMS
Ni et al. [49]	2022	Linear regression, stacking (CatBoost, GBM, LightGBM, RF, SVR, and XGBoost)	Baidu search index, China microdata, epidemic, movie attribute	Total box office performance
Velingkar et al. [50]	2022	CatBoost, LightGBM, RF, voting regression, XGBoost, ridge regression	Budget, cast, crew, genres, IMDb ID, IMDb rating, IMDb vote count, original language, original title, overview, popularity rating, production companies, production countries, release date, revenue, runtime, spoken languages, star power, tagline, TMDb rating, TMDb vote count, title, MPAA rating	Box office revenue

ANOVA, analysis of variance; ARIMA, autoregressive integrated moving average; Bagging, bootstrapped aggregation; CART: classification and regression trees; CatBoost, categorical boosting; DNN, deep neural network; DT, decision tree; eWOM, electronic word-of-mouth; GBM, gradient boosting machine; GPR, gaussian process regression; ISM, independent subspace method; KNN, k-nearest neighbors; LightGBM, light GBM; MLBP, multilayer backpropagation; MLR, multiple linear regression; MGC, mainstream media generated content; MPAA, motion picture association of America; RF, random forest; SVM, support vector machine; SVR, support vector regression; TMDb: The Movie Database; UGC: user-generated content; XGBoost, extreme gradient boosting.

3. Materials and Methods

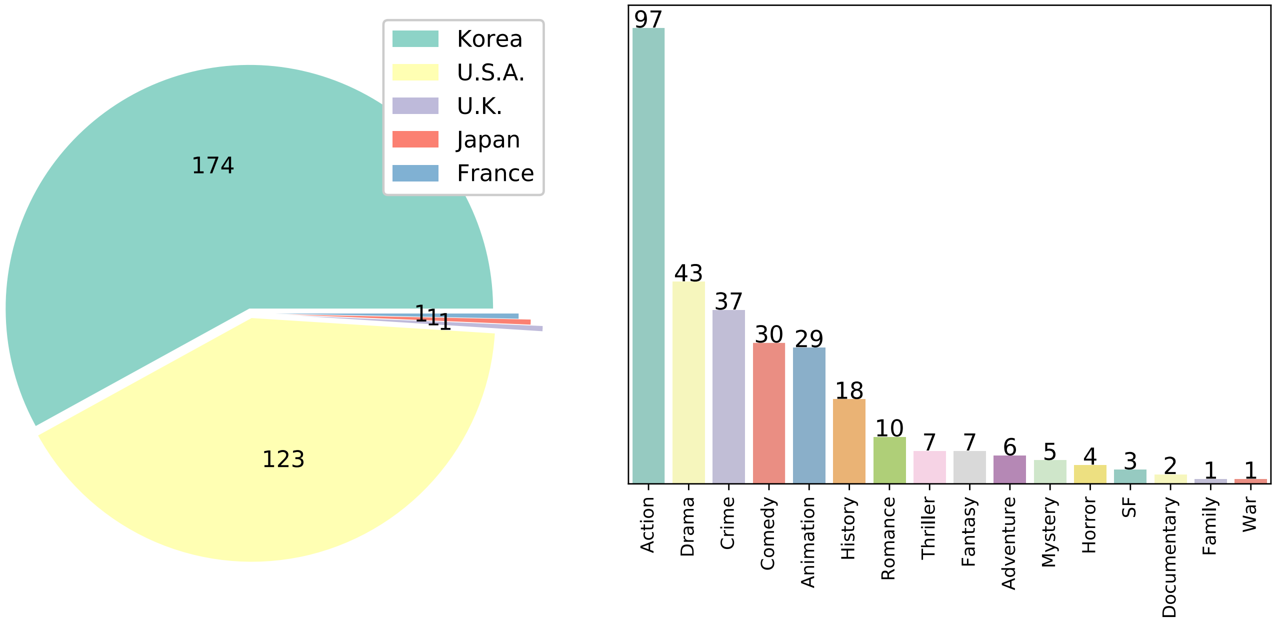
3.1. Data Collection and Preprocessing

We have gathered information about past box office performance through web crawling from the VOD Korea Box-office Information System (VKOBIS) [51], a computer system run by the Korean Film Council. The system quickly and accurately collects and processes movie-viewing data. Our data include the top 300 highest-grossing movies of all time, as shown in Table 2. The information for each movie includes the title, production country, genre, director, actors, release date, and running time. If the data type was a character or category, it was initially collected in Korean and then translated into English by us. This information can be found in Supplementary Materials.

Table 2. Information on the movie dataset.

No.	Name	Description	Data Type
1	Title	Movie title	Character
2	Country	Production country	Category
3	Genre	16 genres	Category
4	Director	Director's name	Character
5	Actor	Leading actor's name	Character

Each movie in the data has only one value for its production country and genre, but it can have zero or more than two values for its director and actors. Figures 1(a) and 1(b) compare the number of movies produced by each country and genre, respectively. Figure 1(a) shows that most movies were produced in South Korea or the US. The only movie made in the United Kingdom (U.K.) is "About Time," the only movie made in Japan is "Your Name," and the only movie made in France is "Taken 2." In Figure 1(b), action movies were the most prevalent, while war and family genres had only one movie each, "Harry Potter and the Sorcerer's Stone" and "Operation Chromite."



(a) Pie chart for country

(b) Bar chart for genre

Figure 1. Information on box office movies.

The title was excluded from the experiment, and the text variables "director" and "actor" were divided into separate values. The values were then labeled "CON" for the country, "GEN" for the genre, "DIRECT" for the director, and "ACT" for the actor. Finally, these character and categorical variables were one-hot encoded, creating dummy variables by marking "1" if the corresponding value was present and "0" if not. The final experimental data had a total of 1028 features.

3.2. DRECE Model Construction

Figure 2 depicts our proposed DRECE (Dimension REduction, Clustering, and Classification for Explainable AI) framework. The framework starts by transforming multidimensional data into 2D data using two stages of dimensionality reduction techniques: deep autoencoder (DAE) and uniform manifold approximation and projection (UMAP). Then, K-means clustering is applied to the reduced data to group similar data points. The clustering result is added as a class label to the data, and one-hot encoding is performed on this class label. This creates variables indicating which movies belong to which cluster. Finally, various machine-learning models, such as logistic regression (LR), DT, RF, and CatBoost, are applied to classify the movie clusters, and the best-performing model is selected. AI techniques that provide insights into the model's decision-making process are used to make the model more explainable.

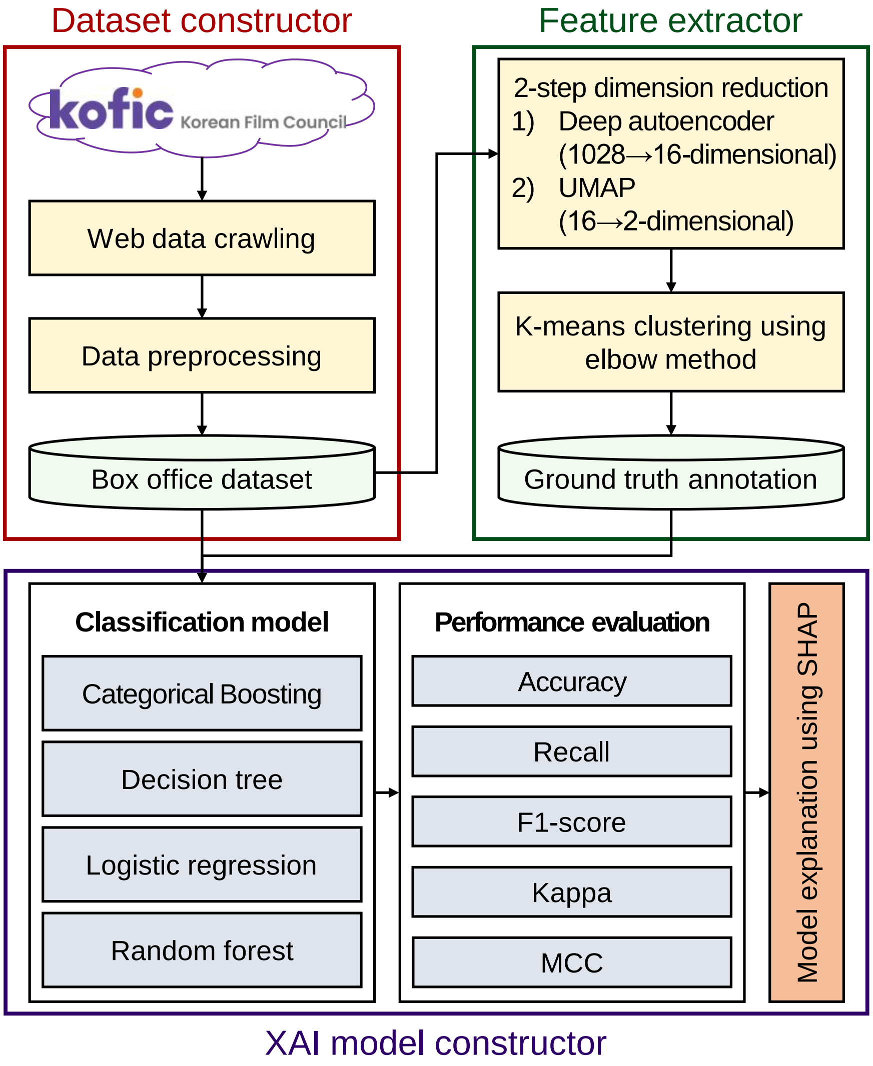


Figure 2. System architecture of the DRECE model.

3.2.1. Dimension Reduction

The reason for dimensionality reduction is that high-dimensional data are challenging to visualize in their raw form and computationally demanding to process [52, 53]. Therefore, by reducing the number of dimensions, the data can be presented in a more easily understandable format, and the computational load is reduced, making processing and analysis faster. As a result, dimensionality reduction can enhance the efficiency, accuracy, and interpretability of data analysis by focusing on the most significant features of the data [53]. We use the DAE's latent variables and UMAP as a means of dimensionality reduction. To visualize a K-means clustering, it must be scaled down to two dimensions.

DAE [54, 55] is a type of NN architecture designed for unsupervised learning. The goal of DAE is to reconstruct the input data. It consists of two main parts: an encoder and a decoder. The encoder transforms the input into a lower-dimensional representation, while the decoder transforms the lower-dimensional representation back into the original data. The idea behind DAE is that it can learn compact representations of data that are more meaningful than the raw input [56]. The complexity of the representation can be increased by adding multiple hidden layers to both the encoder and decoder parts of the network.

The training of DAE involves minimizing the difference between the original input and the output of the decoder. We use mean squared error (MSE) as the loss function for our DAE. The structure of the DAE is shown in Figure 3, and the details on the number of dense layers and the number of units in each layer can be found in the relevant case. We have set the number of units in each layer to decrease or increase by a factor of 2. The activation function for the last layer in the decoder is sigmoid, while we use the rectified linear unit (ReLU) function to activate the other layers.

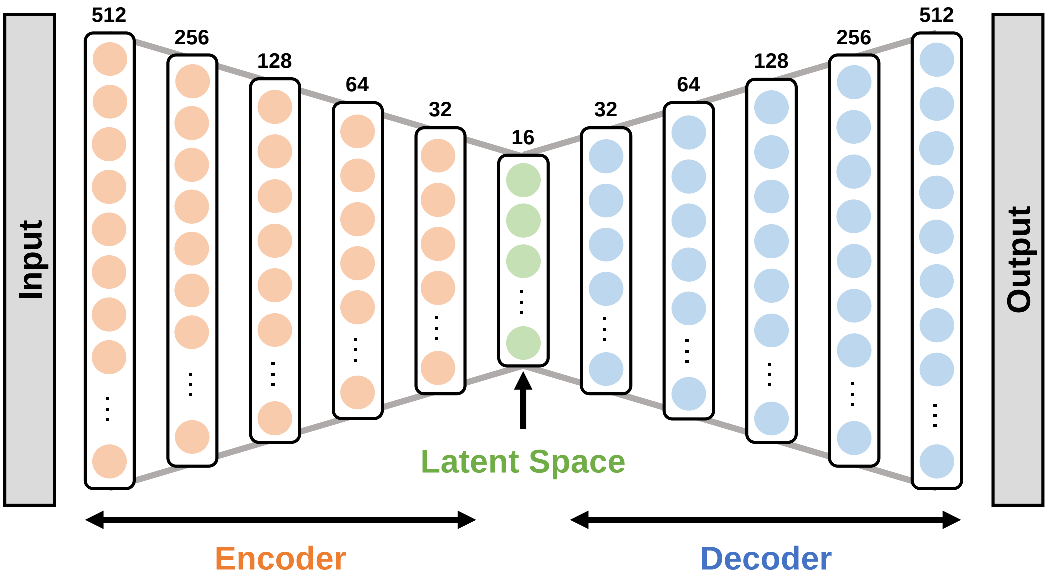


Figure 3. Deep autoencoder architecture.

The latent variable in a DAE is a lower-dimensional representation of the input data created by the encoder. Using latent variables in DAE means that the encoder has learned a compact and meaningful representation of the input data. This low-dimensional representation eliminates fewer essential details and retains the most crucial information about the input. The decoder then utilizes the latent variables to reconstruct the original input, which should be a close approximation of the original data. As a result, latent variables become helpful for data compression, visualization, and feature extraction [54]. In other words, the latent variables of a DAE can be considered concise representations of the input data learned by the network during training [55, 56].

Hence, we extract these latent variables from the DAE and use them. We opt for a manifold learning technique to re-embed the data and aim to learn the entire embedded manifold to optimize the clustering. While the autoencoder we use is a good choice for learning meaningful data representation, it does not consider the local structure [57]. By combining the autoencoder with a manifold learning technique that considers the local structure, we can enhance the quality of the representation in terms of clusterability. Therefore, the dimensionality of the latent space of the DAE is not immediately reduced to two dimensions; instead, a dimensionality reduction process using UMAP is performed separately.

UMAP [58] is a dimensionality reduction algorithm to visualize high-dimensional data in a lower-dimensional space. UMAP combines Riemannian geometry, algebraic topology, and machine learning techniques to find a low-dimensional representation of the data that retains its structure. Unlike dimensionality reduction algorithms like t-distributed stochastic neighbor embedding (t-SNE) [59] and principal component analysis (PCA) [60], UMAP can preserve the data's local and global structure, and the algorithm can be adjusted through various hyperparameters, giving users greater control over the process [57, 58]. Additionally, UMAP is less sensitive to changes in hyperparameters. Due to these features, UMAP is commonly used for data exploration and visualization. To provide a more intuitive and interpretable expression of the clustering results for the preprocessed data, we use DAE and UMAP to reduce the 1030-dimensional data to 2 dimensions (2D).

3.2.2. Clustering Analysis

We use K-means clustering to cluster data points with similar characteristics in 2D data. K-means clustering [28, 61] is a machine learning algorithm that divides a dataset into K clusters. The algorithm repeatedly updates the cluster assignments and cluster centroids until convergence is reached. The number of clusters, K, must be specified beforehand, and the algorithm's goal is to minimize the sum of squares within the clusters. We use the elbow method [32, 62] shown in Figure 4 to determine the optimal value of K. The elbow method involves comparing the sum of squares error (SSE) values for different numbers of clusters by plotting them on a graph. The optimal number of clusters is selected as the number corresponding to the point where the SSE value shows a steep decline followed by a gentle slope.

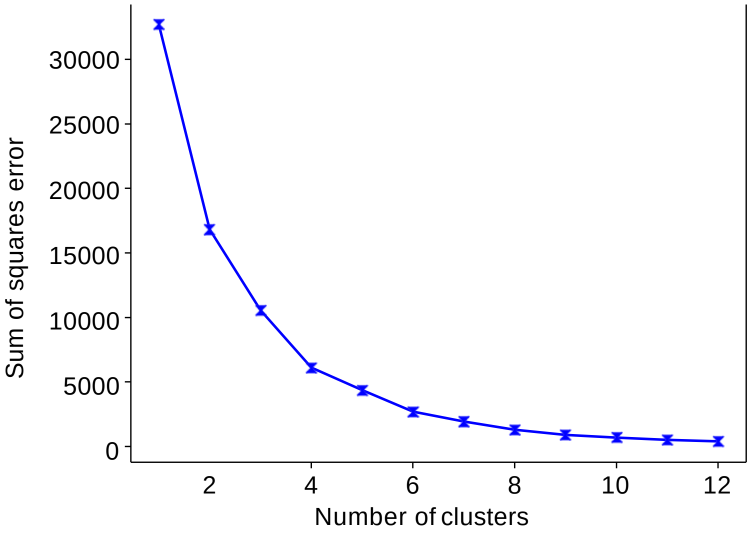


Figure 4. Elbow method for optimal value of K in K-means clustering.

3.2.3. Movie Classification

We create four classification models to effectively present the analysis results of K-means clustering. The original dataset, which consisted of 1030 input variables, posed a challenge as it needed labels or classes. Applying the supervised learning technique to the dataset made it difficult.

To overcome this, we add a new variable by assigning a label of "1" to the box office belonging to each cluster and a label of "0" to the rest. Afterward, we utilize the 1028 variables and the cluster results (i.e., a label of "1" for the data belonging to a specific cluster and a label of "0" for data belonging to other clusters) as input and output variables for each cluster to build four machine learning models, such as LR, DT, RF, and CatBoost. This is because they not only deliver high classification performance when the default values of hyperparameters are set or optimal values suggested in previous studies are used, but they are also straightforward to interpret.

LR [63] is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It uses a logistic function to calculate a particular class's probability, given the independent variables' values. The logistic function generates a value between 0 and 1, representing the likelihood of the dependent variable being a particular class. This probability value is then compared to a threshold, typically 0.5, to classify the dependent variable into one of two categories. LR can be used for binary classification tasks, where the dependent variable can only take two possible values.

DT [45, 64] is a tree-based model used for decision analysis and machine learning. It consists of internal nodes representing tests or conditions on the input features, branches representing the outcomes of these tests, and leaf nodes representing the final prediction. The DT algorithm works by splitting the data into smaller subgroups based on the values of the input features to find the splits that result in the highest accuracy. The final prediction is made based on the values of the input features at the leaf node, and DTs can be used for regression and classification tasks. They help understand the relationships between the input features and the target variable.

RF [50, 65] is an ensemble learning algorithm that uses multiple DTs to make predictions. The algorithm works by training multiple DTs on different samples of data and different subsets of features. Each DT in the RF makes its prediction, and the final prediction is made by taking a majority vote among the predictions made by all trees. This process helps to reduce overfitting, increase accuracy, and make the model more interpretable. By using multiple DTs, RF also reduces the variance of the model, making it more stable. Based on the literature that suggests 128 trees is an appropriate number for RF, we set the number of trees in our model to 128 [23].

CatBoost [23, 66] is an advanced gradient-boosting algorithm that is designed to handle categorical variables effectively. It combines gradient boosting and categorical feature processing techniques to achieve better accuracy than other gradient boosting libraries. This makes it an ideal choice for datasets with many categorical variables. In addition, CatBoost has built-in mechanisms for handling overfitting and missing values, making it more robust and less prone to errors than other gradient-boosting algorithms. CatBoost is a powerful and flexible ensemble learning algorithm well-suited for many data analysis tasks. CatBoost is noted for delivering outstanding performance even with default hyperparameter settings, as noted by the authors, so we proceed with the default values [66].

The above four models, LR, DT, RF, and CatBoost, are designed to be interpretable to provide insight into their decision-making processes. This allows stakeholders to understand how the model arrived at its predictions, which is essential for building trust in the model and making informed decisions based on its outputs. Additionally, these models are easy to set up and require fewer hyperparameters than other complex machine learning models like ensemble learning or deep learning, making them a more accessible and practical choice for many classification tasks. Moreover, RF and CatBoost have the advantage of being easy to interpret and delivering high performance, making them an ideal choice for many classification tasks [23].

3.2.4. Model Interpretation

The original dataset of 1028 input variables needs labeled or classified data, which poses a challenge for implementing supervised learning techniques. As a result, the interpretability of the RF and CatBoost models may be weaker than that of the LR and DT models, which rely on labeled data to make predictions. To address this issue and increase interpretability, we can consider alternative techniques, such as Shapley values, which provide a more in-depth understanding of the model's decision-making process.

Shapley additive explanations (SHAP) [67, 68] is a method that provides explanations for the predictions made by machine learning models. It assigns a contribution value to each feature in the input data to explain the model's predictions. The SHAP values are based on the concept of Shapley values from cooperative game theory and provide a way to fairly represent the contribution of each feature to the prediction, both in absolute terms and relative to other features. SHAP values help interpret complex models and understand the relationship between features and predictions. It helps explain the decisions made by models, identify areas where they may be biased, and provide insight into the decision-making process. By calculating the SHAP values, we can understand which features are essential for predicting movie type, making it a vital reference for understanding the proposed prediction model.

We consider using Tree SHAP for both the RF and CatBoost models to enhance the interpretability of our classification model further. Tree SHAP [23, 39] is a method for interpreting the output of tree-based machine learning models such as RF and CatBoost. It assigns contribution values to each feature in the input data to explain the predictions made by the model. By using Tree SHAP, we can understand how the RF and CatBoost models make decisions and how each feature affects the predictions. This will provide valuable insights into the decision-making process and help identify potential biases in the models. The results of the Tree SHAP analysis will also be used to generate a list of essential features for predicting movie type, which can serve as an essential reference for understanding how the proposed method makes predictions concerning box office type prediction.

4. Results and Discussions

4.1. Movie Data Compression and Clustering Analysis

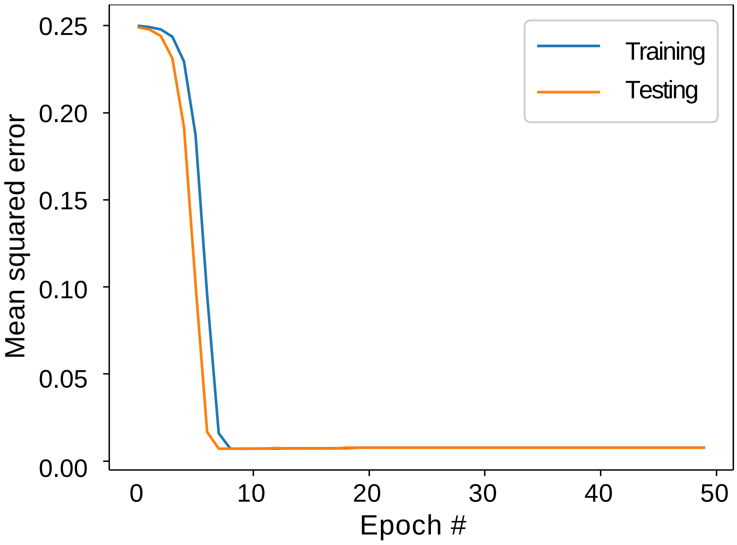


Figure 5. Comparison of training and testing loss for the DAE.

We built a DAE model to condense the 1028 variables (excluding the movie name) from the original 1029 variables into a more compact representation of 16 dimensions. Figure 5 displays the loss of the DAE model during the training and testing phases. The loss was calculated using the MSE method. After 50 training cycles, the loss reached a minimum and stabilized, demonstrating that the DAE model successfully captured the essential features of the data and extracted significant latent variables. Lastly, we utilized the UMAP algorithm to further reduce the 16-dimensional data into a 2-dimensional representation for easier visualization.

Figure 6 displays the results of applying these techniques to the data, showing the results of reducing the data to 2D using both DAE and UMAP, as well as the results of first reducing the data to 16D using DAE and then reducing it to 2D using UMAP. We configured the hyperparameters of UMAP to have a *n_neighbors* value of 8 and a *min_dist* value of 0.2. We found that when the data was reduced to 2D using DAE, only straight lines were displayed, which meant that DAE could not capture the non-linear structure of the data. In contrast, when we used UMAP to reduce the data to 2D, two large distributions were represented along the x-axis, corresponding to the two types of box-office movies we were interested in identifying. However, it was not easy to distinguish between the two types of movies based on the UMAP plot.

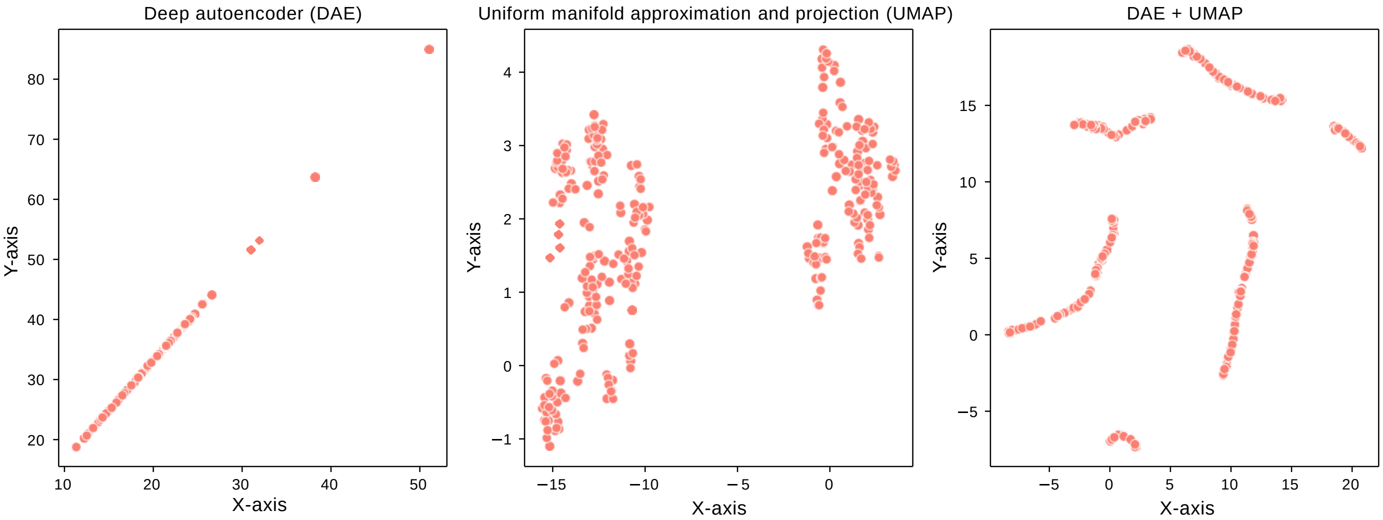
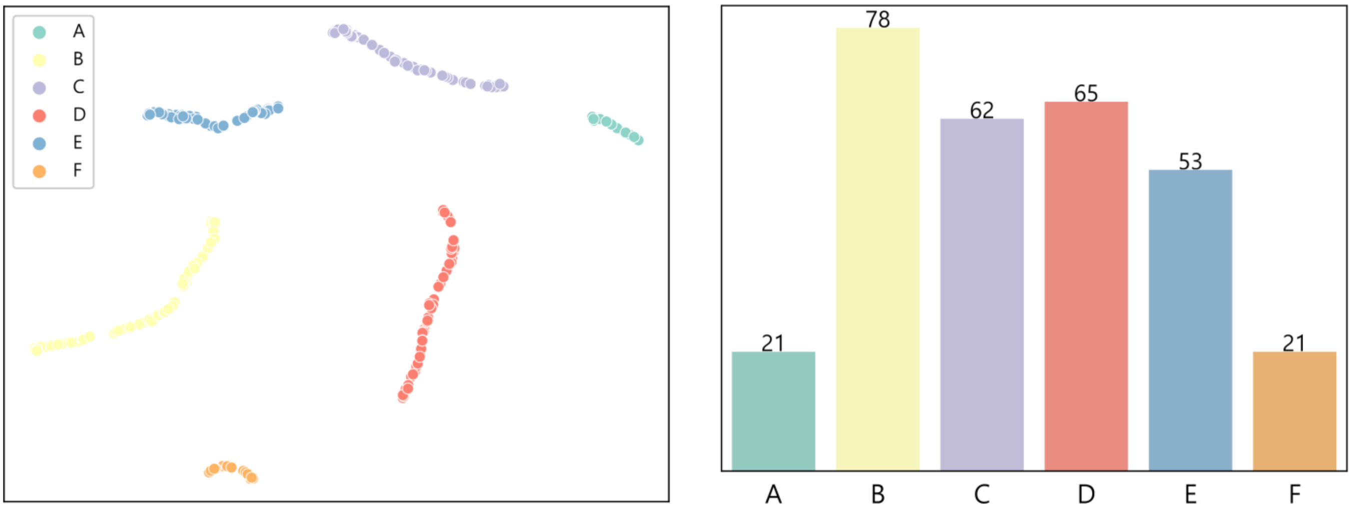


Figure 6. Visualization of dimensionality reduction techniques.

To address this issue, we combined DAE and UMAP techniques by first reducing the data to 16D using DAE and then reducing it to 2D using UMAP. The combination of techniques was more effective in identifying different types of box-office movies. This is because DAE could capture the non-linear structure of the data, while UMAP could preserve both the local and global structure of the data. By using DAE first to reduce the data to a lower dimension and then applying UMAP, we achieved a more accurate and interpretable representation of the data. Hence, we found that combining DAE and UMAP techniques was more effective in identifying different types of box-office movies. The combination of techniques allowed for the capture of the non-linear structure of the data while also preserving the local and global structure of the data.

We used the elbow method to determine the optimal number of clusters. Figure 4 shows the SSE values for different numbers of clusters. We found that reducing the number of clusters from 1 to 6 resulted in a substantial decrease in SSE. However, the decrease in SSE was relatively minor for the number of clusters more significant than 6. Based on these findings, we concluded that this study's optimal number of clusters was 6. The results of the data clustering, using the K-means clustering algorithm on the two-dimensional representation, can be seen in Figure 7(a). Figure 7(b) shows the box office data points distribution across the different clusters.



(a) Result of data clustering into six clusters (b) Number of data points per cluster

Figure 7. Clustering of box office data into six clusters.

4.2. Performance Comaprison of Clssification Models

A confusion matrix, such as the one shown in Table 3, is a table used to evaluate the performance of a binary or multi-class classifier. It provides a clear and concise summary of the classifier's performance by displaying the number of correct and incorrect predictions in a clear format. The rows in the confusion matrix represent the actual class, while the columns represent the predicted class. Each cell in the matrix displays the number of observations that correspond to each combination of actual and predicted classes. In a binary classification confusion matrix, the four most common items are:

- True positives (TP): The number of correctly predicted instances as positive.
- False positives (FP): The number of instances predicted as positive but negative.
- True negatives (TN): The number of correctly predicted instances as negative.
- False negatives (FN): The number of instances predicted as negative but positive.

Table 3. Confusion matrix.

Actual Value	**Predictive Value**	
	Positive	**Negative**
Positive	True positives	False negatives
Negative	False positives	True negatives

When you want to assess how well a classifier works, you can use several metrics calculated using a confusion matrix. The confusion matrix is a table that shows the comparison between the classifier's predicted results and the actual results for a set of data. Using a confusion matrix can also help improve the classifier's performance. To get a full understanding of the classifier's performance, it is best to use the following five metrics together, which are calculated using Equations 1 to 6. A classifier is a tool to assign items to several predefined categories. The performance of a classifier is evaluated by comparing its predicted results to the actual results of a set of data.

- Accuracy: Accuracy is a metric that measures the proportion of correctly classified instances by the classifier. It is calculated by dividing the number of correct classifications by the total number of instances in the data, as described in Equation 1. For example, if a classifier correctly classified 80 out of 100 movie instances, its accuracy would be 80%. However, in cases where the data is imbalanced or false positive or negative results are costly, accuracy may not be the best metric to use. For instance, if a classifier is designed to predict box office hits, and it only correctly identifies 1 out of 10 movies as a hit, its accuracy would still be 10% even though it is not doing an excellent job of predicting box office success.
- Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of box office hits correctly identified by the classifier in the dataset. It is calculated by dividing the number of box office hits correctly identified by the total number of hits in the dataset, as described in Equation 2. For example, if a classifier correctly identified 20 out of 25 box office hits, its recall would be 80%. Recall is an essential metric for evaluating the classifier's ability to identify all box office hits, especially in cases where false negatives are costly. For instance, if a classifier is designed to predict which movies will be box office hits for a film production company, a high recall is crucial to ensure that all potentially successful movies are greenlit for production.
- F1-score: The F1-score is a metric that considers both precision and recall, as described in Equation 3. It provides a balance between precision and recall by considering both metrics at the same time. Precision measures the ratio of the true positive box office hit predictions made by the classifier among all positive predictions. Recall measures the ratio of true positive box office hits among all actual positive box office hits in the data. The F1-score is the harmonic mean of precision and recall and is calculated using a formula. An F1-score of 1 means perfect precision and recall, while 0 means the worst performance.
- Kappa: Cohen's Kappa adjusts for chance agreement and evaluates the agreement between evaluators by considering both the observed agreement and the agreement expected by chance, as described in Equations 4 and 5. Kappa ranges from -1 to 1, with a value of 1 indicating complete agreement between evaluators on the predicted box office hits and actual hits and a value less than 0 indicating agreement worse than chance. In other words, Kappa is a valuable metric for evaluating the performance of a classifier in cases where the data is imbalanced, and it takes into account both the observed agreement and disagreement in the predictions.
- MCC: Matthews correlation coefficient (MCC), as described in Equation 6, measures the quality of a binary classifier used to predict box office types. It takes into account both true and false positive and negative results. MCC ranges from -1 to 1, with a value closer to 1 indicating higher accuracy in the classifier's predictions and a value closer to -1 indicating lower accuracy. If a classifier's predictions are random, its MCC value would be 0. MCC is beneficial in cases where the dataset is uneven, as it considers both accuracy and the ratios of true positive to false positive and true negative to false negative, making it a more reliable measure of performance in these cases.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

$$Recall = TP/(TP + FN) \tag{2}$$

$$F1\text{-}score = TP/(TP + ½ \times (FP + FN)) \tag{3}$$

$$P_e = ((TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + PF))/N^2 \tag{4}$$

$$Kappa = (Accuracy - P_e)/1 - P_e \qquad (5)$$

$$MCC = ((TP \times TN) - (FP \times FN))/\sqrt{}(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)(6)$$

We aimed to perform a feature analysis of the input variables most closely associated with different types of box office movies. To accomplish this, we constructed a classification model, which is a machine-learning model that predicts the class or label of input variables based on their features or characteristics. However, we intended to use something other than this classification model for making future predictions. Instead, we trained the classification model on the entire dataset to identify the input variables that were most strongly related to each type of box-office movie. By doing so, we hoped to gain insights into the key factors that influence the success of different types of movies at the box office.

We used the entire dataset to evaluate the classification model rather than dividing it into training and test sets. This allowed them to assess the model's accuracy on the entire dataset and obtain a more comprehensive understanding of the relationship between input variables and box office movie types. Hence, we constructed a classification model to perform feature analysis on the input variables most closely associated with different types of box office movies. We trained the classification model on the entire dataset to gain insights into the key factors that influence the success of different types of movies at the box office. By using the entire dataset to evaluate the classification model, the authors were able to obtain a more comprehensive understanding of the relationship between input variables and box office movie types.

Table 4. Performance comparison of machine learning models on box office data clusters.

Metric	Model	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F
Accuracy	CatBoost	0.94	0.76	0.83	0.88	0.83	0.94
	DT	0.96	0.78	0.83	0.85	0.87	0.95
	LR	0.98	0.97	0.96	1.00	0.96	0.94
	RF	1.00	1.00	1.00	1.00	1.00	1.00
Recall	CatBoost	0.10	0.06	0.19	0.46	0.04	0.14
	DT	0.38	0.14	0.23	0.34	0.28	0.29
	LR	0.76	0.88	0.79	1.00	0.77	0.19
	RF	1.00	1.00	1.00	1.00	1.00	1.00
F1-score	CatBoost	0.17	0.12	0.32	0.63	0.07	0.25
	DT	0.55	0.25	0.36	0.50	0.43	0.44
	LR	0.86	0.94	0.88	1.00	0.87	0.32
	RF	1.00	1.00	1.00	1.00	1.00	1.00
Kappa	CatBoost	0.16	0.09	0.28	0.57	0.06	0.24
	DT	0.53	0.20	0.30	0.44	0.38	0.43
	LR	0.86	0.92	0.86	1.00	0.85	0.30
	RF	1.00	1.00	1.00	1.00	1.00	1.00
MCC	CatBoost	0.30	0.22	0.40	0.63	0.18	0.37
	DT	0.60	0.33	0.39	0.52	0.47	0.52
	LR	0.87	0.92	0.87	1.00	0.86	0.42
	RF	1.00	1.00	1.00	1.00	1.00	1.00

MCC, Matthews correlation coefficient

Table 4 shows each cluster's performance metrics and the study's machine-learning model. The random state for all models was set to 42, and the decision tree's maximum depth was set to 5 [64]. The random forest model achieved perfect performance for all clusters. However, the logistic regression and decision tree models also performed well, with high accuracy, recall, F1-score, Kappa, and MCC values for many of the clusters. The logistic regression model even outperformed the other models for some clusters, achieving the highest accuracy for clusters A and D, the highest recall for cluster C, and the highest F1-score for clusters A, C, and D. Additionally, the logistic regression model had the highest Kappa and MCC values for clusters A, C, D, and F.

These results suggest that the logistic regression model may be a good choice for identifying specific types of box office movies, particularly for clusters A, C, D, and F. However, it is essential to note that the optimal model choice may vary depending on the specific research question and dataset. Therefore, it is essential to carefully evaluate the performance of different models and consider their strengths and weaknesses before making a final selection. Overall, the results suggest that the random forest model was the most robust and accurate of the models tested, achieving perfect performance for all clusters. The other models performed well for some clusters but showed lower performance for others, suggesting they may need to be more suitable for identifying all box office movie types.

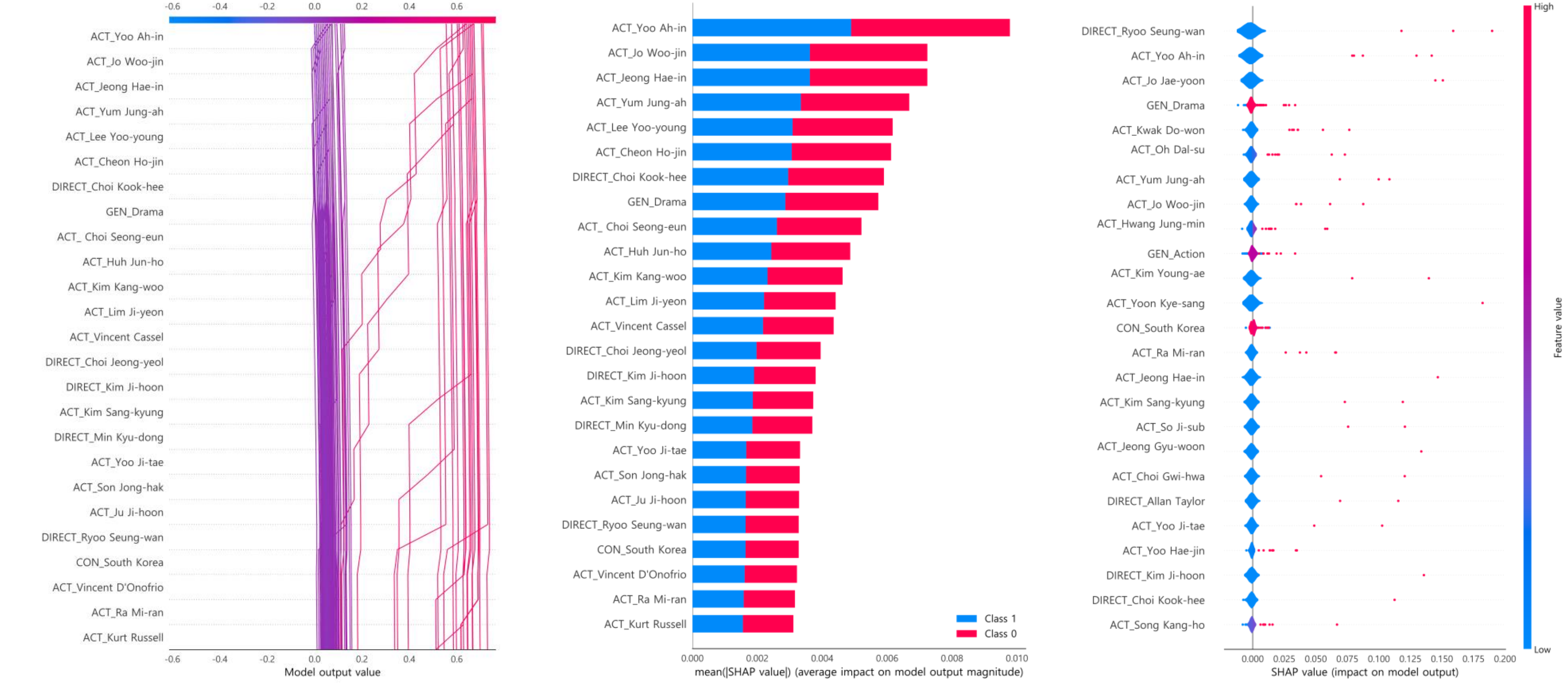
4.3. Interpretability of Box-Office Type Classification Model

Interpretability of the model refers to the ability to understand how it makes its predictions regarding box office-type classification. This is important because it allows us to understand the strengths and weaknesses of the model and how it is likely to perform in real-world scenarios [69]. One way to interpret the model is to examine the model's coefficients for each feature. The coefficients indicate the importance of each feature in the model's predictions for box office-type classification. For example, if a particular director has a high coefficient, the model considers this director to be a strong predictor of the type of box office success the film may have.

The SHAP decision plot visualizes the contribution of each feature to a prediction made by a machine learning model for a specific instance or data point. The plot shows the relationship between features and the prediction, with the essential features considered first. Features with higher absolute SHAP values have a more significant impact on the prediction. The plot displays the prediction as the mean prediction of the leaf nodes, with the leaf nodes' size indicating each feature's contribution. Interactions between features are represented by branches in the plot, with the branch's size indicating the feature's effect on the prediction. The value of the feature shows how much it contributes compared to the average contribution of all features for all instances. The SHAP decision plot provides an interpretable explanation of the relationship between features and a model's prediction.

The SHAP summary plot is a visual representation of the contribution of each feature to the prediction of a binary classification model. The plot displays the SHAP values for each feature, representing the feature's impact on the model's prediction. The x-axis displays the features, while the y-axis represents the contribution of each feature to the prediction, with positive values indicating a positive contribution and negative values indicating a negative contribution. The SHAP summary plot provides a clear view of the relationship between features and the model's prediction, making it an effective tool for understanding the behavior of binary classification models, such as movie genre classification or box office revenue prediction.

The SHAP violin plot is a graphical representation of the distribution of SHAP values for each feature in a binary classification model. The x-axis displays the features, and the y-axis shows the SHAP values split into two halves, one for the positive class and one for the negative class. The width of the plot at a point on the y-axis represents the density of the SHAP values for that feature, with the median being the center. The position of the plot on the y-axis indicates the direction of the feature's contribution to the prediction, with positive contributions to the right of the baseline and negative contributions to the left. The width of the plot provides information on the variance in contributions, with more expansive plots indicating higher variance and narrower plots indicating lower variance.

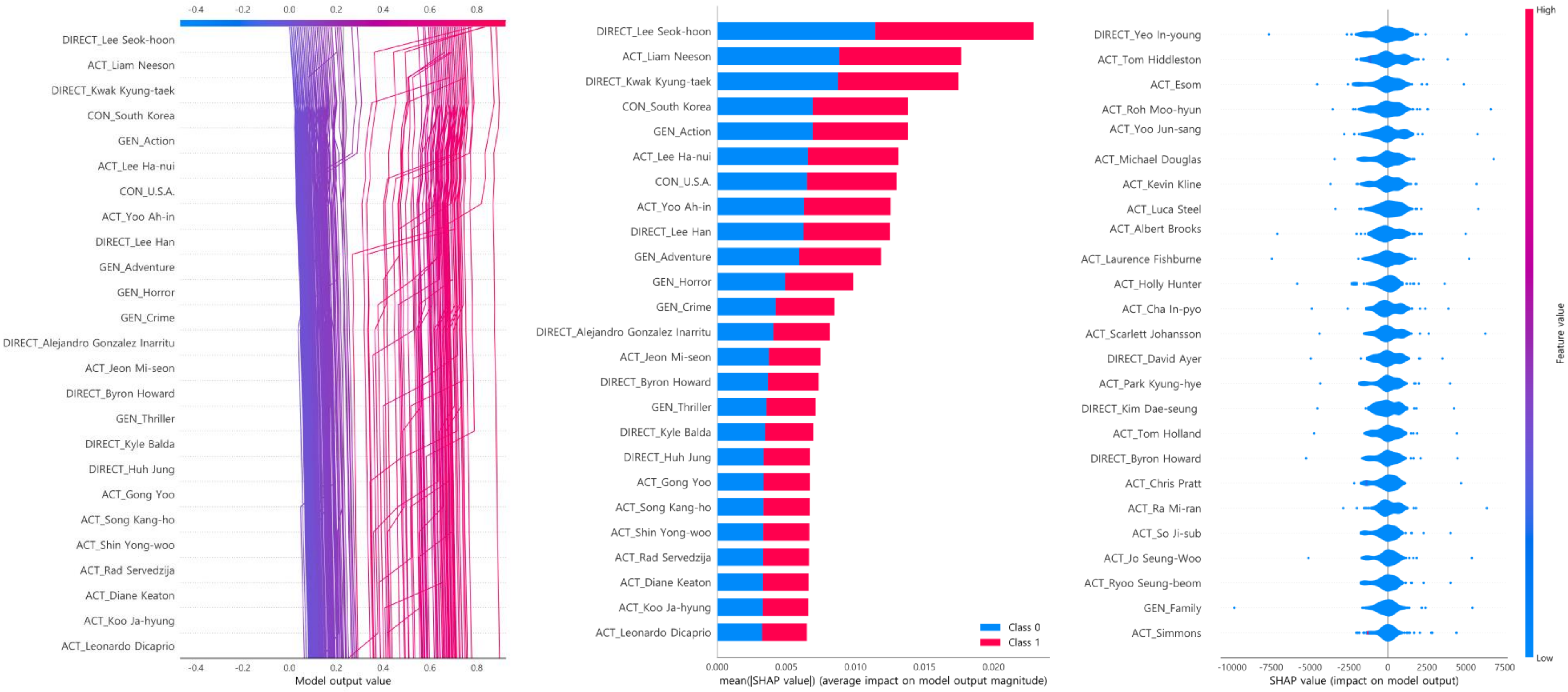


(a) Decision plot

(b) Summary plot

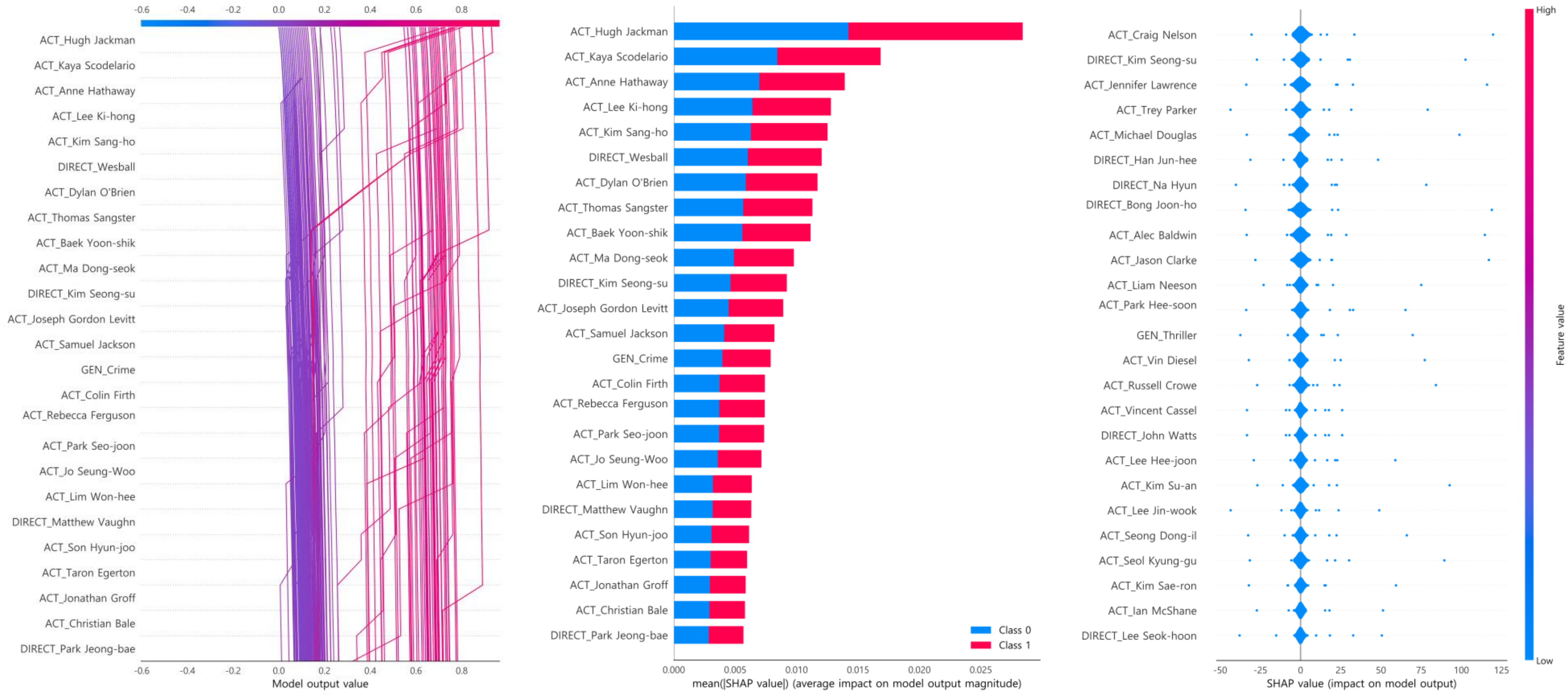
(c) Violin plot

Figure 8. Shapley additive explanations (SHAP) plots for Cluster A.



(a) Decision plot (b) Summary plot (c) Violin plot

Figure 9. SHAP plots for Cluster B.

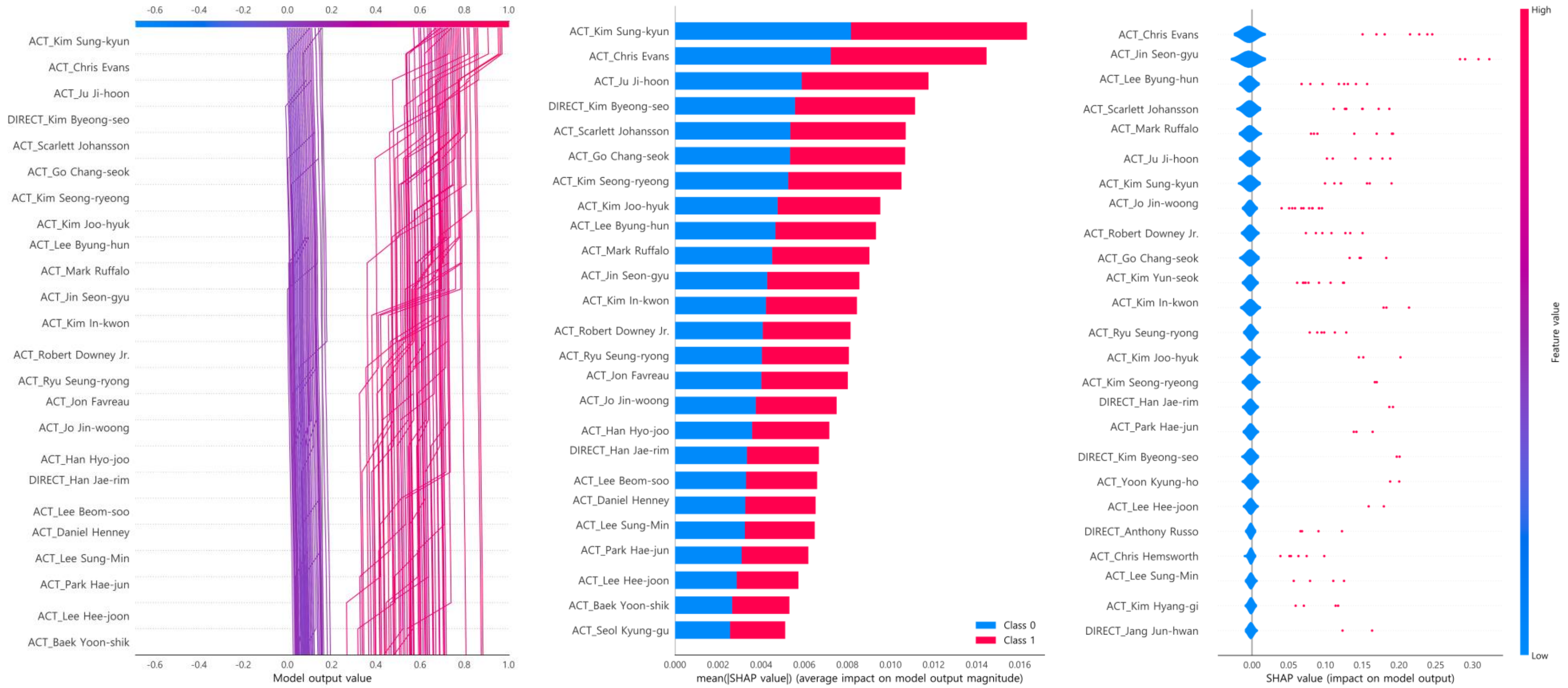


(a) Decision plot

(b) Summary plot

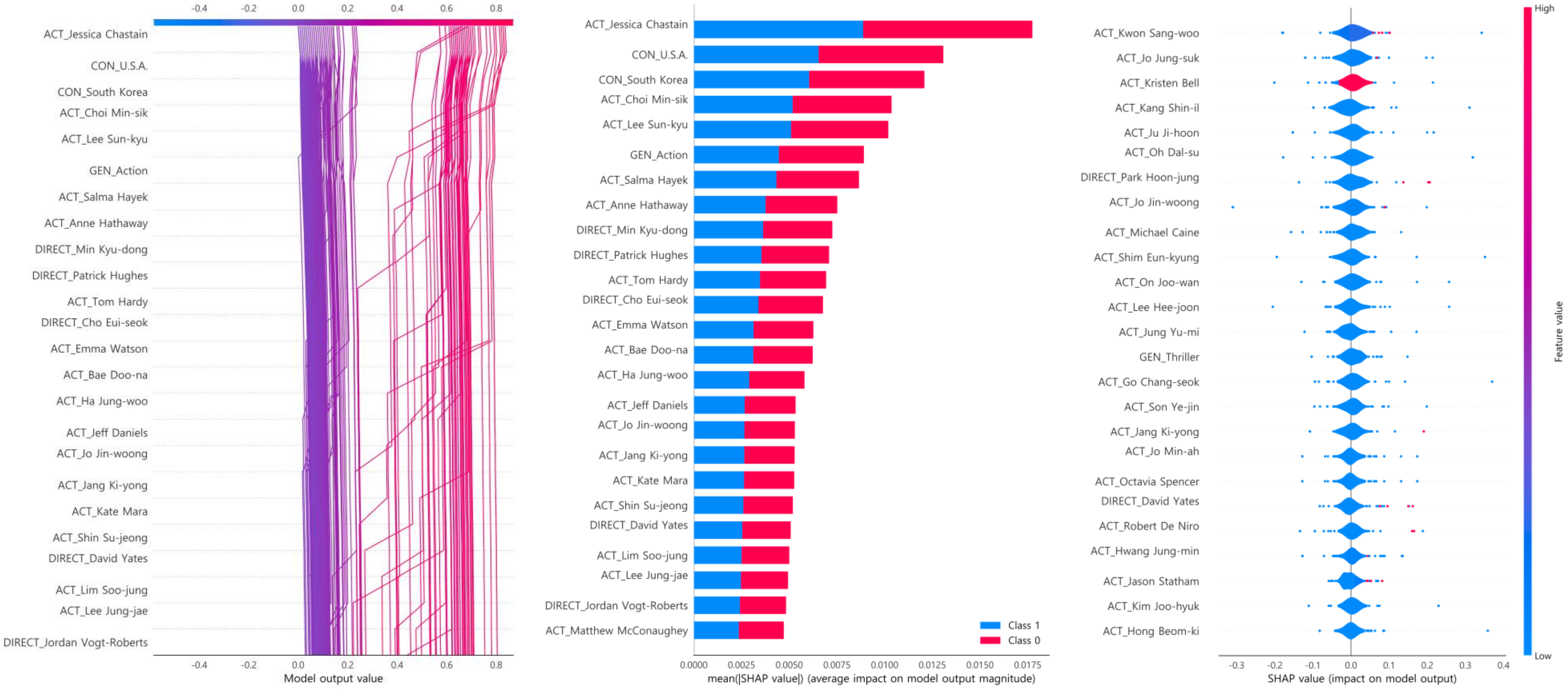
(c) Violin plot

Figure 10. SHAP plots for Cluster C.



(a) Decision plot (b) Summary plot (c) Violin plot

Figure 11. SHAP plots for Cluster D.

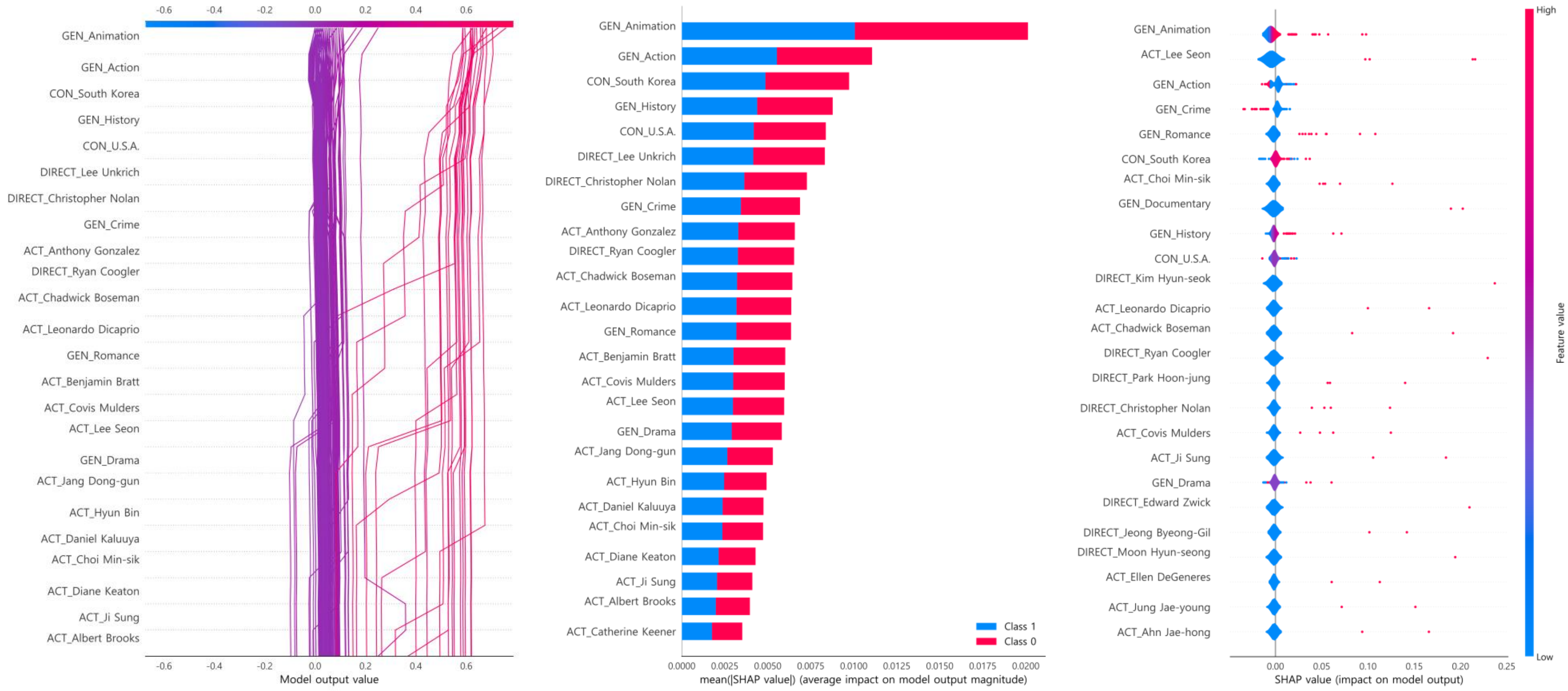


(a) Decision plot

(b) Summary plot

(c) Violin plot

Figure 12. SHAP plots for Cluster E.



(a) Decision plot (b) Summary plot (c) Violin plot

Figure 13. SHAP plots for Cluster F.

As shown in Figure 8, we analyzed Cluster A and revealed that it contains keywords such as "Yoo Ah-in" and "National Bankruptcy Day." It is found that Ah-in Yoo, a famous Korean actor, is the most influential variable in this cluster. Furthermore, the other actors and directors from "Day of National Bankruptcy," a movie where Yoo Ah-in played the main character, are also part of this cluster. Along with this movie, the cluster consists of other directors or actors who have appeared alongside Yoo Ah-in in different movies. Interestingly, the French actor Vincent Cassel is also included in this cluster, despite having yet to appear in a Korean movie. The article suggests that if Vincent Cassel and other non-Korean actors were to work with a Korean actor or director, it could lead to a good synergy effect. Combining various actors' charms and acting skills, as well as reflecting cultural diversity, can result in a more affluent and exciting work.This could enhance the work's perfection and attract the audience's interest.

In Cluster B, shown in Figure 9, the main keywords are "Taken" and "animation." The "Taken" series is a popular action movie with three installments, and the actors in the series are critical in this cluster. Liam Neeson, who played the main character in all three movies, is the most influential actor in the cluster. Many Korean directors, such as Lee Seok-hoon and Kwak Gyeong-taek, have produced successful films in Korea and are highly regarded in the industry. The cluster also contains many directors and voice actors who have produced American animations. If they participate in animation works produced in Korea, it can increase the possibility of gaining popularity worldwide. Since action movies and animations are distributed together in this cluster, new works can be created by producing action movies as animations or live-action animations. These initiatives can open new markets and create new revenue opportunities. Overall, this cluster provides insight into potential collaborations between Korean and American actors, directors, and animators that could result in successful and globally famous works.

Cluster C is characterized by movies with suspenseful and action-packed plots, such as the famous "Maze Runner" and "Kingsman" series, as shown in Figure 10. These movies are known to captivate viewers with their thrilling action scenes and suspenseful developments, and the influence of the actors and directors who made them possible is significant. Notable Korean actors in this cluster include Kim Sang-ho, Baek Yun-sik, and Ma Dong-seok, known for their impressive performances in villainous roles, particularly in action scenes. Their inclusion in the cluster suggests the potential for them to be cast as the protagonists of new action movies within this cluster. In addition, Cluster C also includes actors with exceptional singing abilities, such as Hugh Jackman and Cho Seung-woo. When combined with suitable directors or staff, this cluster shows promise for involvement in film production within the musical genre. Given the variety of styles within Cluster C, there is great potential for creating various types of action movies. However, new attempts and challenges will be required to achieve this, and these efforts could open new markets for the industry.

Cluster D is mainly influenced by middle-aged male Korean actors and actors who have appeared in the Avengers series, as shown in Figure 11. Notable actors in this cluster include Kim Seong-gyun, Ju Ji-hoon, Ko Chang-seok, and Lee Byung-hun. They could create a Korean version of the Avengers with their combined influence. Additionally, Korean actresses such as Kim Seong-ryeong and Han Hyo-joo have a significant presence in this cluster. If they were to take on roles like those played by Scarlett Johansson in The Avengers, they could attract public attention with their acting skills and charm. Furthermore, the actors in this cluster have the potential to participate in international activities, like the Avengers actors. As Korean actors become more involved in international film production, they could help elevate the Korean film industry to a global level. This could lead to a more diverse and globally recognized film industry in Korea.

As depicted in Figure 12, Cluster E comprises actors who have shown their acting skills in films of various genres, including "Interstellar" and "Harry Potter," which are representative works containing fantasy elements. The actors in this cluster are characterized by taking on diverse roles and displaying a wide range of acting skills, not limited to a single genre. Among the actors in this cluster, Jessica Chastain stands out as the most influential actress due to her outstanding performance in "Interstellar." She has established herself as a performer who can perform in films of various genres and is highly recognized for her acting skills. As a result, the E cluster has the potential to create films of various genres by combining the appropriate actors. Recently, Jang Ki-yong and Lim Soo-jeong, actors from Cluster E, were selected for lead roles in a drama. These actors have shown consistent acting abilities in a variety of works and are expected to receive high marks in dramas.Additionally, since there are many actors with acting ability and charm regardless of genre in this cluster, it is expected that actors receiving attention from various works will appear in the future.

As shown in Figure 13, Cluster F primarily comprises actors and directors who have contributed to the success of famous animated films such as Coco, Finding Dory, and Pororo. These movies have gained worldwide recognition and popularity, and it is believed that the technical abilities of the directors, the voice actors' acting skills, and the use of visual effects all played essential roles in their success. On the other hand, Korean actors belonging to this cluster, such as Jang Dong-gun and Hyun Bin, are known for their versatility and are active in various fields, including advertising, dramas, and movies. Specifically, Dong-gun Jang, Bin Hyun, Min-shik Choi, and Ji-seong all share a history of appearing in historical drama films. Given the strengths of these actors, it is recommended that they appear as voice actors in animated movies or movies that heavily utilize computer-generated (CG) technology. This is because these actors can showcase their acting skills while also taking advantage of the impressive visual effects that CG technology can produce. Following these recommendations, we can expect Korean actors to create movies featuring innovative acting performances and stunning visual effects.

4.4. Discussion

Although CatBoost is known for its excellent performance even with default hyperparameter values, the results in Table 4 did not meet our expectations. As a result, we decided to perform a grid search to find the optimal hyperparameters for CatBoost, including the learning rate, maximum depth of each tree, and the coefficient at the L2 regularization term of the cost function [70]. After specifying a wide range of hyperparameter values and running the grid search, the researchers achieved perfect performance for all clusters, with accuracy, recall, F1-score, Kappa, and MCC equal to 1. However, it is worth noting that while this approach led to perfect performance, it comes with a high computing cost. Specifying a wide range of hyperparameter values and running a grid search can be computationally intensive and time-consuming, especially for large datasets. Therefore, this approach may only sometimes be feasible or practical in real-world applications, and trade-offs between performance and computational efficiency must be considered.

The impact of COVID-19 on recent movie box office performance could not be studied in this study because the films did not perform well and were thus excluded from box office data. Korean and Hollywood films were also excluded and could not be analyzed. Regrettably, a comprehensive analysis of movie reviews on Korean portals such as Naver or KaKao was not performed. Differences in box office performance between successful and unsuccessful films were not examined. The research aimed to create a classification model for box office types to aid decision-makers in maximizing profits in the Korean cinema industry. The top 300 box office datasets from the Korean Film Council's online database were analyzed using machine learning and data mining techniques. Factors affecting a movie's box office success in South Korea were identified using input variables. Our model provides valuable insights for decision-makers in the film industry to make data-driven decisions and improve future film success in the Korean market.

Additionally, this research focuses primarily on the Korean film market, and there is regret that a more comprehensive and integrated analysis of the global film industry could not be performed. More advanced natural language processing techniques will likely be utilized to provide a more systematic analysis of these topics. The plan is to build an integrated platform that covers the worldwide film industry. As a result, it is recommended to leave these topics for future research, which will include a planned analysis of Hollywood data as well.

5. Conclusions

Data-driven decision-making models utilize data analysis to inform decision-making in the film industry. These models analyze various factors related to a film's potential success, such as box office revenue and audience size, to predict a film's performance be-fore its release. Filmmakers and investors can then use this information to reduce risk and make more informed decisions about the film's production, marketing, and distribution. Data-driven decision-making models can potentially revolutionize the film industry by providing valuable insights and improving efficiency. By accurately predicting a film's success, these models can help filmmakers and investors make better decisions about the resources they allocate toward producing and promoting a film. This, in turn, can lead to better-performing films and a more sustainable film industry overall. Predicting the success of a film can be challenging due to the complexity of the film industry and the various factors that impact a film's success, such as cultural and economic systems. However, data-driven decision-making models have the potential to help address these challenges and provide more accurate predictions and better decision-making support for filmmakers and investors in the film industry.

To gain a better understanding of the factors that contribute to the success of box-office movies, we analyzed the top 300 highest-grossing movies of all time. The data was collected using web crawling from the VKOBIS, a computer system run by the Korean Film Council, and included information on the title, production country, genre, director, actors, release date, and running time of each movie. We used the DRECE framework to process this data, which involves transforming multi-dimensional data into 2D data through dimensionality reduction techniques such as DAE and UMAP. The 2D data were then subjected to K-means clustering to group similar data points and classify the movie clusters. Finally, we applied machine-learning models, including LR, DT, RF, and CatBoost, to classify the movie clusters. The results showed that the RF model performed best, with accuracy and recall of 1.00 and an F1-score, Kappa, and MCC of 1.00. Our findings provide valuable insights into the factors that influence the success of box-office movies and can inform future filmmaking decisions.

Our DRECE framework analysis has uncovered a world of possibilities for the Korean film industry. Our research has identified six unique clusters, each with strengths and characteristics. Cluster A, featuring the influence of Korean actor Yoo Ah-in and French actor Vincent Cassel, highlights the potential for captivating collaborations between Korean and non-Korean actors. Cluster B, focusing on action movies and animations, suggests exciting opportunities for Korean and American actors, directors, and animators to join forces. Cluster C, dominated by thrilling and action-packed films, presents a chance for Korean actors to shine in villainous roles and for actors with exceptional singing abilities to showcase their talents. Cluster D, influenced by Korean actors and actresses, hints at creating a Korean version of the Avengers and the potential for Korean actors to take the world by storm. Cluster E, known for its diverse acting skills, holds promise for various films across different genres. Finally, Cluster F, characterized by animated films and CG technology, opens the door for Korean actors to showcase their acting skills through voice acting and visually stunning works. Our findings offer many insights into the Korean film industry's potential for success and innovation.

Unfortunately, the impact of COVID-19 on recent movie box office performance could not be studied, and Korean and Hollywood films were excluded from the analysis. Furthermore, a comprehensive analysis of movie reviews on Korean portals and the differences in box-office performance between successful and unsuccessful films were not examined. However, our model provides valuable insights for decision-makers in the film industry to make data-driven decisions and improve future film success in the Korean market. Future research is necessary to thoroughly analyze the impact of COVID-19 on the box-office performance of recent movies and to perform a more comprehensive and integrated analysis of the global film industry. Advanced natural language processing techniques will likely be utilized to provide a more systematic analysis, and the plan is to build an integrated platform that covers the worldwide film industry. As a result, it is recommended to leave these topics for future research, which will include a planned analysis of Hollywood data as well.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1.

Author Contributions: Conceptualization, S.L., J.O., and D.S.; methodology, S.L.; software, S.L.; validation, J.O. and D.S.; formal analysis, D.S.; investigation, J.O.; resources, J.M.; data curation, S.L. and J.M.; writing—original draft preparation, S.L. and J.O.; writing—review and editing, J.M.; visualization, D.S.; supervision, J.M.; project administration, J.M.; funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Soonchunhyang University Research Fund.

Data Availability Statement: The data presented in this study are available in the Supplementary Materials.

Acknowledgments: We sincerely appreciate the support provided by BK21 FOUR (Fostering Outstanding Universities for Research) (No. 5199990514663).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gul, R.; Leong, K.; Mubashar, A.; Al-Faryan, M.A.S.; Sung, A. The Empirical Nexus between Data-Driven Decision-Making and Productivity: Evidence from Pakistan's Banking Sector. *Cogent Business & Management* **2023**, *10*, 2178290.
2. Jafari, M.; Ahmadi Safa, M. Data use in language schools: The case of EFL teachers' data-driven decision making. *Journal of Educational Change* **2022**, 1-22.
3. Nouinou, H.; Asadollahi-Yazdi, E.; Baret, I.; Nguyen, N.Q.; Terzi, M.; Ouazene, Y.; Yalaoui, F.; Kelly, R. Decision-making in the context of Industry 4.0: Evidence from the textile and clothing industry. *Journal of Cleaner Production* **2023**, 136184.
4. Maiti, M.; Vuković, D.; Mukherjee, A.; Paikarao, P.D.; Yadav, J.K. Advanced data integration in banking, financial, and insurance software in the age of COVID-19. *Software: Practice and Experience* **2022**, *52*, 887-903.
5. Yang, J.; Xiu, P.; Sun, L.; Ying, L.; Muthu, B. Social media data analytics for business decision making system to competitive analysis. *Information Processing & Management* **2022**, *59*, 102751.
6. Chen, L.; Liu, H.; Zhou, Z.; Chen, M.; Chen, Y. IT-business alignment, big data analytics capability, and strategic decision-making: Moderating roles of event criticality and disruption of COVID-19. *Decision Support Systems* **2022**, *161*, 113745.
7. Zizic, M.C.; Mladineo, M.; Gjeldum, N.; Celent, L. From industry 4.0 towards industry 5.0: A review and analysis of paradigm shift for the people, organization and technology. *Energies* **2022**, *15*, 5221.
8. Kondapaka, P.; Khanra, S.; Malik, A.; Kagzi, M.; Hemachandran, K. Finding a fit between CXO's experience and AI usage in CXO decision-making: evidence from knowledge-intensive professional service firms. *Journal of Service Theory and Practice* **2023**.
9. Maja, M.M.; Letaba, P. Towards a data-driven technology roadmap for the bank of the future: Exploring big data analytics to support technology roadmapping. *Social Sciences & Humanities Open* **2022**, *6*, 100270.
10. Teng, Y.; Zhang, J.; Sun, T. Data-driven decision-making model based on artificial intelligence in higher education system of colleges and universities. *Expert Systems* **2022**, e12820.
11. Mollá, N.; Heavin, C.; Rabasa, A. Data-driven decision making: New opportunities for DSS in data stream contexts. *Journal of Decision Systems* **2022**, *31*, 255-269.
12. Kalsoom, A.; Maqsood, M.; Yasmin, S.; Bukhari, M.; Shin, Z.; Rho, S. A computer-aided diagnostic system for liver tumor detection using modified U-Net architecture. *The Journal of Supercomputing* **2022**, *78*, 9668-9690.
13. Jabeen, A.; Yasir, M.; Ansari, Y.; Yasmin, S.; Moon, J.; Rho, S. An Empirical Study of Macroeconomic Factors and Stock Returns in the Context of Economic Uncertainty News Sentiment Using Machine Learning. *Complexity* **2022**, *2022*.
14. Maqsood, H.; Maqsood, M.; Yasmin, S.; Mehmood, I.; Moon, J.; Rho, S. Analyzing the stock exchange markets of EU nations: a case study of brexit social media sentiment. *Systems* **2022**, *10*, 24.
15. Li, X.; Ding, Y. Holistic governance for sustainable public services: Reshaping government–enterprise relationships in China's digital government context. *International Journal of Environmental Research and Public Health* **2020**, *17*, 1778.
16. Yang, Z.; Liu, C.; Nie, R.; Zhang, W.; Zhang, L.; Zhang, Z.; Li, W.; Liu, G.; Dai, X.; Zhang, D. Research on Uncertainty of Landslide Susceptibility Prediction—Bibliometrics and Knowledge Graph Analysis. *Remote Sensing* **2022**, *14*, 3879.

17. Zhang, Y.; Yan, Q. Landslide susceptibility prediction based on high-trust non-landslide point selection. *ISPRS International Journal of Geo-Information* **2022**, *11*, 398.
18. Pourhashem, G.; Malichová, E.; Piscová, T.; Kováčiková, T. Gender Difference in Perception of Value of Travel Time and Travel Mode Choice Behavior in Eight European Countries. *Sustainability* **2022**, *14*, 10426.
19. Xie, Y.; Chen, Y.; Lian, Q.; Yin, H.; Peng, J.; Sheng, M.; Wang, Y. Enhancing real-time prediction of effluent water quality of wastewater treatment plant based on improved feedforward neural network coupled with optimization algorithm. *Water* **2022**, *14*, 1053.
20. Kogut, I.; Armbruster, F.; Polak, D.; Kaur, S.; Hussy, S.; Thiem, T.; Gerhardts, A.; Szwast, M. Antibacterial, Antifungal, and Antibiotic Adsorption Properties of Graphene-Modified Nonwoven Materials for Application in Wastewater Treatment Plants. *Processes* **2022**, *10*, 2051.
21. Ansari, Y.; Yasmin, S.; Naz, S.; Zaffar, H.; Ali, Z.; Moon, J.; Rho, S. A Deep Reinforcement Learning-Based Decision Support System for Automated Stock Market Trading. *IEEE Access* **2022**, *10*, 127469-127501.
22. Maqsood, M.; Yasmin, S.; Gillani, S.; Aadil, F.; Mehmood, I.; Rho, S.; Yeo, S.-S. An autonomous decision-making framework for gait recognition systems against adversarial attack using reinforcement learning. *ISA transactions* **2022**.
23. Moon, J.; Rho, S.; Baik, S.W. Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with Shapley values. *Sustainable Energy Technologies and Assessments* **2022**, *54*, 102888.
24. Nikolic, D.; Kostic-Stankovic, M.; Jeremic, V. Market Segmentation in the Film Industry Based on Genre Preference: The Case of Millennials. *Engineering Economics* **2022**, *33*, 215-228.
25. Zhang, H.; Wu, Y. The Analysis and Implementation of Film Decision-Making Based on Python. *Scientific Programming* **2022**, *2022*.
26. Gemignani, Z.; Gemignani, C.; Galentino, R.; Schuermann, P. *Data fluency: Empowering your organization with effective data communication*; John Wiley & Sons: 2014.
27. Loy, J. Project-Based Supply Chain Intelligence and Digital Fabrication for a Sustainable Film Industry. In *Supply Chain Intelligence: Application and Optimization*, Springer: 2020; pp. 37-59.
28. Mbunge, E.; Fashoto, S.G.; Bimha, H. Prediction of box-office success: A review of trends and machine learning computational models. *International Journal of Business Intelligence and Data Mining* **2022**, *20*, 192-207.
29. Lipizzi, C.; Iandoli, L.; Marquez, J.E.R. Combining structure, content and meaning in online social networks: The analysis of public's early reaction in social media to newly launched movies. *Technological Forecasting and Social Change* **2016**, *109*, 35-49.
30. Baek, H.; Oh, S.; Yang, H.-D.; Ahn, J. Electronic word-of-mouth, box office revenue and social media. *Electronic Commerce Research and Applications* **2017**, *22*, 13-23.
31. Basnayake, H.; Jayalal, S. Predicting box office success of movies using sentiment analysis and opinion mining. In Proceedings of the International Research Symposium on Pure and Applied Sciences (IRSPAS 2016), Faculty of Science, University of Kelaniya, Sri Lanka, 2016; p 88.
32. Darban, Z.Z.; Valipour, M.H. GHRS: Graph-based hybrid recommendation system with application to movie recommendation. *Expert Systems with Applications* **2022**, *200*, 116850.
33. Ding, C.; Cheng, H.K.; Duan, Y.; Jin, Y. The power of the "like" button: The impact of social media on box office. *Decision Support Systems* **2017**, *94*, 77-84.
34. Panaligan, R.; Chen, A. Quantifying movie magic with google search. *Google Whitepaper—Industry Perspectives+ User Insights* **2013**.
35. Mestyán, M.; Yasseri, T.; Kertész, J. Early prediction of movie box office success based on Wikipedia activity big data. *PloS one* **2013**, *8*, e71226.
36. Chon, W. Beyond the International Film Festival: Contact Zones for the Agonistics and Solidarity. In *Korean Film and Festivals*, Routledge: 2023; pp. 81-97.
37. Parc, J. Evaluating the effects of protectionism on the film industry: A case study analysis of Korea. *Handbook of State Aid for Film: Finance, Industries and Regulation* **2018**, 349-366.
38. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18.
39. Lee, J.; Jeong, J.; Jung, S.; Moon, J.; Rho, S. Verification of De-Identification Techniques for Personal Information Using Tree-Based Methods with Shapley Values. *J. Pers. Med*. **2022**, *12*, 190.
40. Kim, K.H. *Hegemonic mimicry: Korean popular culture of the twenty-first century*; Duke University Press: 2021.
41. Zhang, L.; Luo, J.; Yang, S. Forecasting box office revenue of movies with BP neural network. *Expert Systems with Applications* **2009**, *36*, 6580-6587.
42. Kim, T.; Hong, J.; Kang, P. Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting* **2015**, *31*, 364-390.
43. Hur, M.; Kang, P.; Cho, S. Box-office forecasting based on sentiments of movie reviews and independent subspace method. *Information Sciences* **2016**, *372*, 608-624.
44. Lee, S.; Bikash, K.; Choeh, J.Y. Comparing performance of ensemble methods in predicting movie box office revenue. *Heliyon* **2020**, *6*, e04260.
45. Lee, S.; Choeh, J.Y. Movie production efficiency moderating between online word-of-mouth and subsequent box office revenue. *Sustainability* **2020**, *12*, 6602.

46. Bogaert, M.; Ballings, M.; Van den Poel, D.; Oztekin, A. Box office sales and social media: a cross-platform comparison of predictive ability and mechanisms. *Decision Support Systems* **2021**, *147*, 113517.
47. Pan, C. Research on the Influencing Factors of Box Office: A Case Study of the Top 100 Films in the Chinese Box Office in 2019. In Proceedings of 2021 5th Annual International Conference on Data Science and Business Analytics (ICDSBA); pp. 189-195.
48. Li, D.; Liu, Z.-P. Predicting Box-Office Markets with Machine Learning Methods. *Entropy* **2022**, *24*, 711.
49. Ni, Y.; Dong, F.; Zou, M.; Li, W. Movie Box Office Prediction Based on Multi-Model Ensembles. *Information* **2022**, *13*, 299.
50. Velingkar, G.; Varadarajan, R.; Lanka, S. Movie Box-Office Success Prediction Using Machine Learning. In Proceedings of 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T); pp. 1-6.
51. VKOBIS, Available online: https://www.vkobis.or.kr/boxoffice/selectBoxofficeHistoryList.do (accessed on 25 February 2023)
52. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records. *Informatics* **2020**, *7*, 17.
53. Tang, Y.B.; Chen, D.; Li, X.L. Dimensionality reduction methods for brain imaging data analysis. *ACM Comput. Surv.* **2021**, *54*, 87.
54. González-Muñiz, A.; Díaz, I.; Cuadrado, A.A.; García-Pérez, D. Health indicator for machine condition monitoring built in the latent space of a deep autoencoder. *Reliability Engineering & System Safety* **2022**, *224*, 108482.
55. Kim, J.-Y.; Cho, S.-B. Explainable prediction of electric energy demand using a deep autoencoder with interpretable latent space. *Expert Systems with Applications* **2021**, *186*, 115842.
56. Darban, Z.Z.; Valipour, M.H. GHRS: Graph-based hybrid recommendation system with application to movie recommendation. *Expert Systems with Applications* **2022**, *200*, 116850.
57. An, J.; Ai, P.; Liu, C.; Xu, S.; Liu, D. Deep clustering bearing fault diagnosis method based on local manifold learning of an autoencoded embedding. *IEEE Access* **2021**, *9*, 30154-30168.
58. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint* arXiv:1802.03426 2018.
59. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
60. Maćkiewicz, A.; Ratajczak, W. Principal components analysis (PCA). *Computers & Geosciences* **1993**, *19*, 303-342.
61. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* **1979**, *28*, 100-108.
62. Bholowalia, P.; Kumar, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications* **2014**, *105*.
63. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic regression*; Springer: 2002.
64. Verwer, S.; Zhang, Y. Learning decision trees with flexible constraints and objectives using integer optimization. In Proceedings of Integration of AI and OR Techniques in Constraint Programming: 14th International Conference, CPAIOR 2017, Padua, Italy, June 5-8, 2017, Proceedings 14; pp. 94-103.
65. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R news* **2002**, *2*, 18-22.
66. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* **2018**, *31*.
67. Le, T.-T.-H.; Kim, H.; Kang, H.; Kim, H. Classification and explanation for intrusion detection system based on ensemble trees and SHAP method. *Sensors* **2022**, *22*, 1154.
68. Kim, M.; Kim, D.; Jin, D.; Kim, G. Application of Explainable Artificial Intelligence (XAI) in Urban Growth Modeling: A Case Study of Seoul Metropolitan Area, Korea. *Land* **2023**, *12*, 420.
69. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. XAI—Explainable artificial intelligence. *Science robotics* **2019**, *4*, eaay7120.
70. grid_search – CatBoost. Available online: https://catboost.ai/en/docs/concepts/python-reference_catboost_grid_search (accessed on 25 February 2023).