Towards an Effective Over-The-Top Platform Service: A Machine Learning Approach for Box Office Analysis

Subeen Leem
Department of Medical Science
Soonchunhyang University
Asan, South Korea
qlsl0519@sch.ac.kr

Jisong Oh
Department of AI and Big Data
Soonchunhyang University
Asan, South Korea
jso2562@sch.ac.kr

Jihoon Moon
Department of AI and Big Data
Soonchunhyang University
Asan, South Korea
jmoon22@sch.ac.kr

***Abstract*—We developed a machine learning approach for trend analysis of box office performance on online platforms. We first collected the top 300 box office data points, including movie name, release date, production country, genre, director, actor, and screening time from the Korean Film Council online integrated computer network. We preprocessed these data to configure a total of 1030 input variables and then used a deep autoencoder and uniform manifold approximation and projection to reduce them to two input variables. Next, we used a k-means clustering algorithm to cluster movies with similar characteristics on the reduced data. We constructed an extreme gradient boosting (XGBoost) model using the 1030 original variables and cluster results as input and output variables for each cluster. We finally demonstrated the interpretability of the XGBoost model using Shapley additive explanations to understand the clustering results. The proposed method is expected to assist in the marketing strategy of domestic entertainment companies because it can be used for casting for movie production, determining release time, and adjusting running time by determining elements that determine box office hits.**

Keywords—box office analysis, data reduction, data clustering, ensemble learning, explainable artificial intelligence

I. INTRODUCTION

In the present time, an increasing number of people seek entertainment through watching movies. For the steady growth of the film industry, it is essential to systematically analyze various movie box office factors before making movies [1]. An effective box office prediction model can provide business decision support and guidance for film production and distribution companies, which is of immense significance for the sustainable development of the film industry.

Machine learning (ML) is a research hotspot in artificial intelligence (AI), and it has steadily been used in a multitude of life activities. Many researchers have used various ML techniques to analyze the factors that influence movie box office success [2–4]. Furthermore, numerous experts have investigated the impact of the COVID-19 pandemic on the box office, which is currently experiencing a slump. For example, Shen et al., [5] concluded that the pandemic has had a significant adverse effect on the worldwide box office.

As studies on movie box office prediction [2–4] have traditionally concentrated chiefly on constructing ML models to predict audience numbers, studies analyzing the factors that impact the movie box office, especially on online platforms, are relatively rare.

In this paper, we propose an ML approach for trend analysis of box office on online platforms in South Korea. To do so, we use the top 300 box office data points collected from the Korean Film Council (KOFIC) online integrated computer network.

The main contributions of this paper are as follows:

- We configure various input variables representing movie characteristics to determine the factors influencing a movie's box office performance in South Korea.
- We reduce the feature dimension to avoid dimensional curses for effective clustering of movies and then perform box office trend analysis according to each cluster's characteristics.
- We use an explainable AI (XAI) technique to interpret the factors affecting the box office. It helps to set diverse targets for box office improvement by movie types.

The rest of the paper is organized as follows: Section II describes the proposed approach including data preprocessing, dimensionality reduction, clustering analysis, and interpretable model. Section III presents experimental settings and results. Section IV concludes and outlines future research directions.

II. MATEIALS AND METHODS

A. Data Preprocessing

This study collected box office data of the top 300 movies, including movie name, release date, production country, genre, director, actor, and screening time by web crawling on the KOFIC online integrated computer network.

Only "month" was considered for the release date to reflect the characteristics of the release season. Each box office consists of five production countries—South Korea, United States, United Kingdom, Japan, and France—and 16 genres—science fiction (SF), family, horror, documentary, drama, melo/romance, mystery, crime, historical drama, thriller, animation, action, adventure, war, comedy, and fantasy. Figures 1 and 2 compare and show the number of works for the country or genre of production, respectively.

For the production country, genre, director, and actor variables, we generated a total of 1026 dummy variables through one-hot encoding by marking each movie as "1" if it corresponds to the variable and "0" otherwise. Finally, a total of 1030 variables represent the box office factors that influenced a movie's box office performance, including movie name, month, screening time, and these dummy variables.

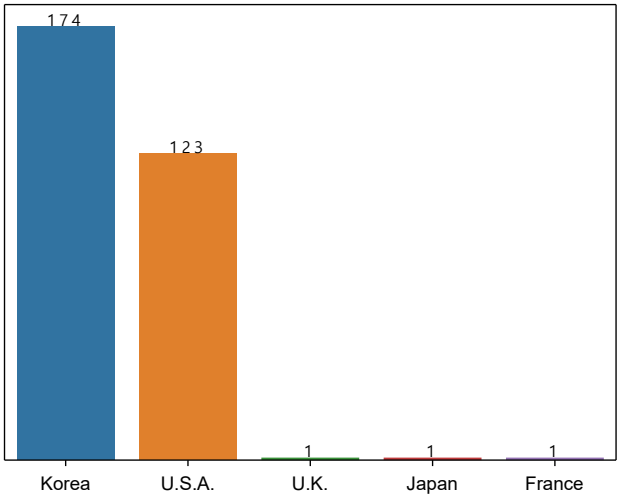

Fig. 1. Number of movies by country.

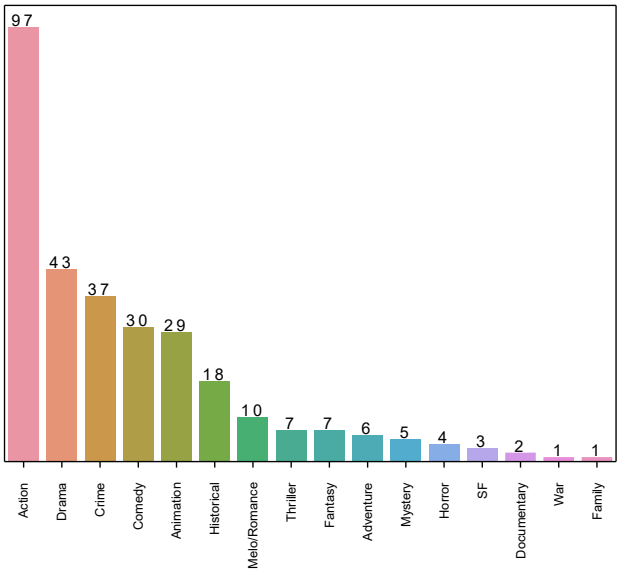

Fig. 2. Number of movies by genre.

B. *Dimensionality Reduction*

1) DAE: A deep autoencoder (DAE) [6] is an artificial neural network trained through unsupervised learning. The structure of a DAE consists of an encoder, a latent space, and a decoder. It encodes data, and then the encoded data are decoded as close as possible to the input data. In this method, we extract latent variables because they can be compressed through the encoder to adequately explain the input data.

2) UMAP: An uniform manifold approximation and projection (UMAP) [7] is a manifold learning technique for dimension reduction. The UMAP performs well in terms of visualization quality and preserves many global structures with superior runtime performance. Furthermore, it has no computational restrictions on the embedding dimension, making it viable as a general-purpose dimension reduction technique for machine learning.

C. *Clustering Analysis*

We cluster reduced data with similar characteristics using a k-means clustering algorithm [8]. K-means clustering works by minimizing the variance of the distance difference between each cluster and clustering the given data into *K* clusters. We consider the elbow method [9] to detect the optimal number of clusters. The elbow method compares the sum of square error (SSE) value according to the number of clusters through a graph to select the number of clusters corresponding to the part where the SSE value shows a sharp slope and then a gentle slope.

D. *Interpretable Model*

Ensemble learning techniques strategically generate and combine multiple models, resulting in better performance than a single model. These techniques derive excellent performance from a tabular dataset [10–13]. For each cluster clustered in the previous step, we use an extreme gradient boosting (XGBoost) algorithm, one of the popular ensemble learning techniques, to construct the interpretable models for box office performance.

The XGBoost algorithm [11, 12] has been optimized for efficiency, flexibility, and portability. This algorithm uses the gradient descent algorithm and adapts a parallel tree-boosting strategy that improves accuracy and speed while building a model. The XGBoost algorithm boosts decision trees by correcting earlier mistakes to improve future performance.

III. Experimental Results

We constructed a DAE model to reduce the 1029 variables (excluding the movie name from the 1030 original variables) to 16-dimensional data. Figure 3 shows the loss value for epochs while training and testing the DAEs we implement. The loss is set to mean square error (MSE). After 50 learning cycles, the loss is minimal and stable, implying that this model captures the features of data well and extracts meaningful latent variables from the data. We then reduced the data from 16 dimensions to 2 dimensions using UMAP to be able to visualize the data in a plane.

As a result of the elbow method's experiment, as shown in Fig. 4, we confirmed that the number of clusters from 1 to 4

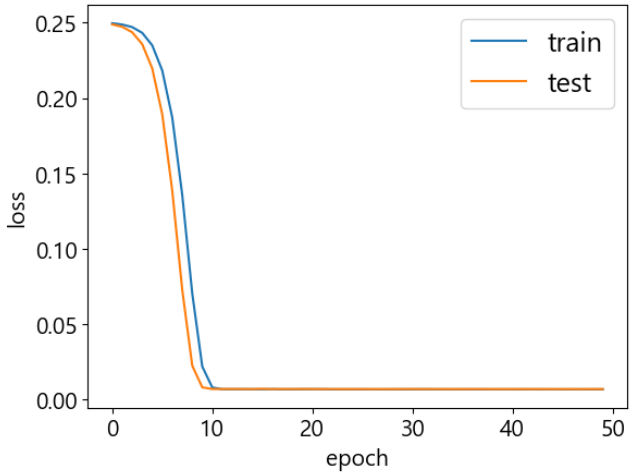

Fig. 3. Loss according to epoch for training and testing sets.

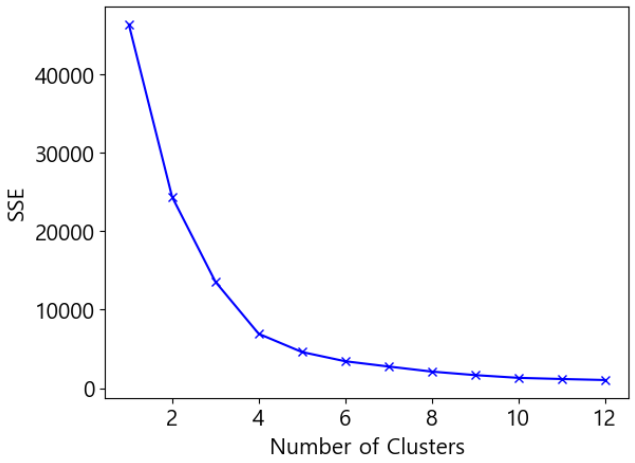

Fig. 4. Elbow method for finding optimal number of clusters.

decreased by a large margin, and then the SSE decreased by a small margin from the number of clusters above 5. Therefore, the optimal K was set to "4" in this study. Figure 5 shows the results of clustering data reduced to two dimensions using a k-means clustering algorithm, and Fig. 6 presents the number of box office data points per cluster.

Finally, we constructed an XGBoost model using the 1030 original variables and cluster results (i.e., "1" for the corresponding clustering data and "0" for the other clustering data) as input and output variables for each cluster. We then applied Shapley additive explanations (SHAP) [10, 12] to the XGBoost model for the determination of the factors that influenced a movie's box office performance. The SHAP is a tool that can effectively correlate independent and dependent variables in model learning based on the Shapley value.

Visualization for model analysis was performed from Figs. 7 to 10. These figures show the importance of the top 7 variables for independent variables that have a substantial influence on the composition of the prediction model through SHAP values. The X-axis represents the SHAP value, while the Y-axis represents the independent variables. Each cluster represents a set of highly successful movies having certain features in common. On all clusters, movie screen time was the most important in predicting box offices in all clusters, and release months were always included in the top 7.

1) Cluster A: The importance of variables for Cluster A is shown in Fig. 7. The time and the month of release have the most impact. The high importance of director Chris Buck and director Andrew Stanton, who make lots of animated films, can be observed. In addition, we confirmed that the upper variables are English-American directors.

2) Cluster B: The importance of variables for Cluster B is shown in Fig. 8. Actors Oh Dal-soo, Ma Dong-seok, and Shim Eun-kyung are of high importance. Relatively, Korean actors are of high importance.

3) Cluster C: The importance of variables for Cluster C is shown in Fig. 9. Actors Robert Downey Jr., Song Kang-ho, and Lee Jung-jae show high importance, and we confirmed that this cluster consist of works featuring middle-aged men.

4) Cluster D: The importance of variables for Cluster D is shown in Fig. 10. Ryu Seung-ryong, a prolific actor, was the most important in this cluster.

IV. Conclusion

We developed a machine learning approach for trend analysis of box office performance on online platforms. We used DAE, UMAP, K-means clustering, and an ensemble model to predict online movie box office success. First, we collected box office data for the top 300 movies through web crawling and performed data preprocessing for input variable configuration. Next, we reduced the input variables from 1029 dimensions to two dimensions through DAE and UMAP. Then, we clustered the reduced data into four types using K-means clustering. Finally, we constructed an XGBoost model trained by clusters and interpreted the results of the XGBoost model through the SHAP bar plots for each cluster.

Through an experiment, the presence of English American animation movies first implies a high likelihood of box office success for movies falling in Cluster A. The presence of Korean actors then implies a high likelihood of box office

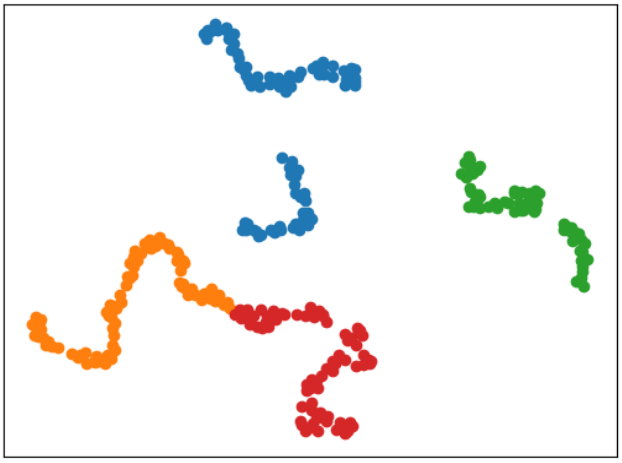

Fig. 5. Result of UMAP with 2-dimensional data.

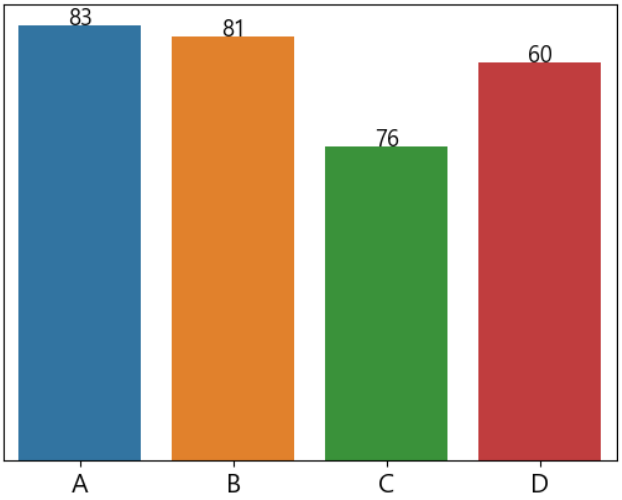

Fig. 6. Number of data per cluster.

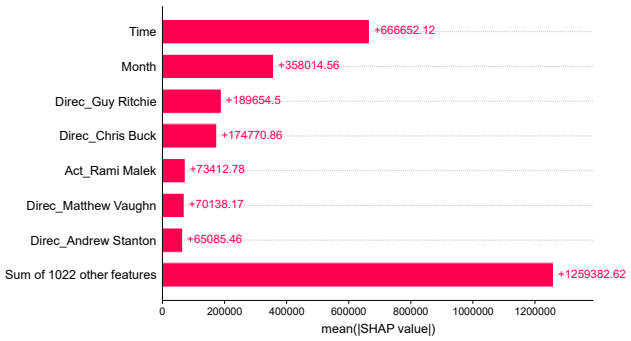

Fig. 7. SHAP bar plot for Cluster A.

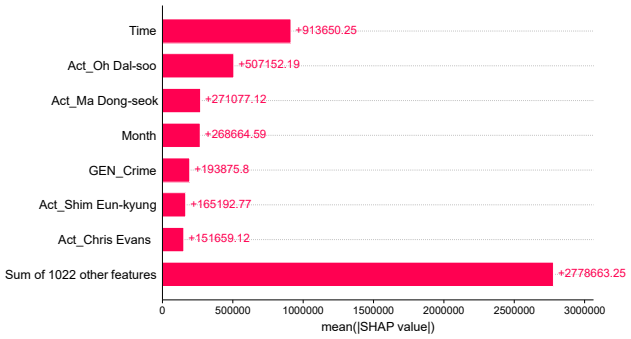

Fig. 8. SHAP bar plot for Cluster B.

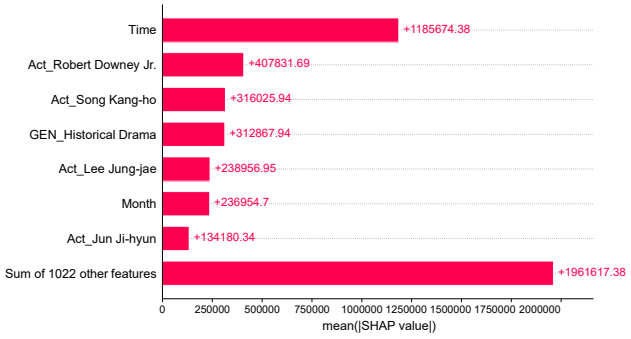

Fig. 9. SHAP bar plot for Cluster C.

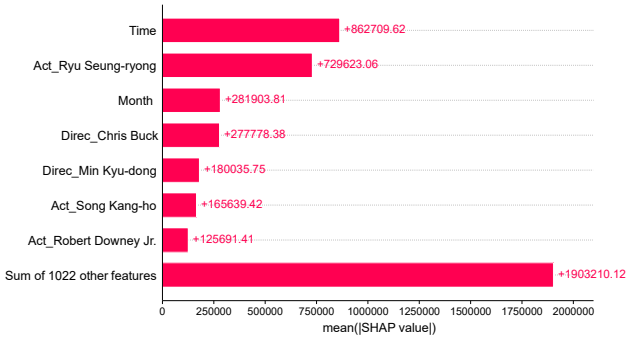

Fig. 10. SHAP bar plot for Cluster D.

success for movies falling in Cluster B. After that, the presence of middle-aged male actors implies a high likelihood of box office success for movies falling in Cluster C. Finally, the presence of Ryu Seung-ryong implies a high likelihood of box office success for movies falling in Cluster D.

In the future, we plan to further analyze the impact of the movie industry by adding various movie data other than the top 300 and COVID-19-related variables, as well as the classification of general or independent/art movies.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation) in 2021 (2021-0-01399) and the Soonchunhyang University Research Fund (No. 20221183).

REFERENCES

[1] S. Y. Kim, S. H. Im, and Y. S. Jung, “A Comparison Study of the Determinants of Performance of Motion Pictures: Art Film vs. Commercial Film,” *The Journal of the Korea Contents Association*, vol. 10, no. 2, pp. 381–393, 2010.

[2] D. Li and Z.-P. Liu, “Predicting Box-Office Markets with Machine Learning Methods,” *Entropy*, vol. 24, no. 5, p. 711, 2022.

[3] J. Zhao, F. Xiong, and P. Jin, “Enhancing Short-Term Sales Prediction with Microblogs: A Case Study of the Movie Box Office,” *Future Internet*, vol 14, no. 5, p. 141, 2022.

[4] H. Shen, M. Fu, H. Pan, Z. Yu, and Y. Chen, “The Impact of the COVID-19 Pandemic on Firm Performance,” *Emerging Markets Finance and Trade*, vol. 56, no. 10, pp. 2213–2230, 2020.

[5] T. Kim, J. Hong, and P. Kang, “Box Office Forecasting considering Competitive Environment and Word-of-Mouth in Social Networks: A Case Study of Korean Film Market,” *Computational Intelligence and Neuroscience*, vol. 2017, p. 4315419, 2017.

[6] M. Son, J. Moon, S. Jung, and E. Hwang, “A Short-Term Load Forecasting Scheme Based on Auto-Encoder and Random Forest,” in *Proc. the 3rd International Conference on Applied Physics, System Science and Computers (APSAC)*, 2018, pp. 138–144.

[7] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[8] Y. J. Jeong, J. Lee, J. Moon, J. H. Shin, and W. D. Lu, “K-means Data Clustering with Memristor Networks,” *Nano Letters*, vol. 18, no. 7, pp. 4447–4453, 2018.

[9] P. Bholowalia and A. Kumar, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” *International Journal of Computer Applications*, vol. 105, no. 9, 2014.

[10] J. Lee, J. Jeong, S. Jung, J. Moon, and S. Rho, "Verification of De-Identification Techniques for Personal Information Using Tree-Based Methods with Shapley Values," *Journal of Personalized Medicine*, vol. 12, no. 2, p. 190, 2022.

[11] J. Moon, S. Rho, and S. W. Baik, "Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with Shapley values," *Sustainable Energy Technologies and Assessments*, vol. 54, p. 102888, 2022.

[12] S. Park, J. Moon, S. Jung, S. Rho, S. W. Baik, and E. Hwang, "A Two-Stage Industrial Load Forecasting Scheme for Day-Ahead Combined Cooling, Heating and Power Scheduling," *Energies*, vol. 13, no. 2, p. 443, 2020.

[13] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.