A movie box office revenues prediction algorithm based on human-machine collaboration feature processing

DOI : 10.36909/jer.ICCSCT.19489

Dongqi Wang[1], Yanqing Wu[2], Chenmin Gu[3], Yiqin Wang[3], Xingyu Zhu[3], Weihua Zhou[1], Xin(Maxwell) Lin[1]*

[1]Department of Data Science and Management Engineering, School of Management, Zhejiang University, Hangzhou, 310058, China.
[2]Meta-intelligence and Decision Lab., Hangzhou OR Cloud Co.ltd, Hangzhou, 311121, China.
[3]Ningbo High School, Ningbo, 315199, China.

* Corresponding Author: linx612@zju.edu.cn

ABSTRACT

Improving the accuracy of box office revenue forecasts is conducive to stimulating the creation, market investment, infrastructure construction, and rational allocation of public resources in the film market, as well as promoting social welfare and cultural prosperity. Since the existing box office revenue prediction algorithm does not consider film industry structure, the prediction accuracy is not satisfying. This paper firstly builds a two-stage human-machine collaborative feature processing framework. In the first stage, based on the box office data, the regression decision tree algorithm is used to process all the box office features preliminarily and delete the unimportant features automatically. In the second stage, feature processing is coupled with the built Artificial Neural Network (ANN). In this stage, the features processed by the machine are manually classified, and multiple, incompatible feature sets are divided. After designing the incompatible set network pruning algorithm, the neural network is pruned. We construct the data set with a total of 7098 movies crawled online on four platforms. Numerical experimental results show that the Mean Absolute Error (MAE) of the two-stage algorithm is significantly better than the baseline model, which can effectively reduce the noise caused by encoding between incompatible features directly, improve the prediction accuracy of ANN, accelerate the forward inference speed of ANN and reduce the consumption of computing resources.

Keywords: Box Office Revenue Prediction; Feature Engineering; Decision Tree; Artificial Neural Network; Neural Network Pruning; Human-Machine Collaboration.

INTRODUCTION

The film industry has been the focus of economists and the general public since the last century. The film box office prediction is an issue that has been continuously focused on by the investment community. The film industry has been developing continuously for over 100 years in the United States. From 2010 to 2019, the film industry maintained a box office revenue of more than 10 billion yuan yearly. Investing in the life cycle of a film, such as film preparation, casting, shooting, pre-release, scheduling, and other stages, plays a crucial role. Once there is a reliable box office prediction function, it will provide a good source of information for each investment from a micro perspective. From a macro perspective, it will allow funds to flow to more commercially promising producers, thus optimizing the ecology of the overall film market. The global epidemic has greatly affected the film market in recent years. In 2020, the box office income tax in the United States was reduced to $2 billion, down 80.7% from 2019. In such an environment, investment risk is greater, and accurate box office forecasts are more important than ever.

In traditional box office prediction, manual analysis is one of the more common methods. People conduct model or non-model analysis based on several movie features, such as budget, director, MPAA score, etc. At present, with the development of artificial intelligence, more and more people are paying attention to the use of machine learning methods to predict the box office of theaters (Liu et al., 2019), which has achieved better prediction results than the former because of the efficient use of data by machine learning. However, these learning-based methods only perform statistical analysis based on data. They do not consider the prior knowledge of human beings on box office predictions, so their predictive effects are still limited. In real life, it is not uncommon for the deduction results of machine learning algorithms to conflict with the actual business execution process because people's prior knowledge is not considered, such as the fully automatic driving AI algorithm, whose safety has been questioned (Kiran et al., 2021), partial reinforcement learning models that are hard to converged in the training process (Wang et al., 2019), "Meituan rider" trapped in the algorithm, etc.

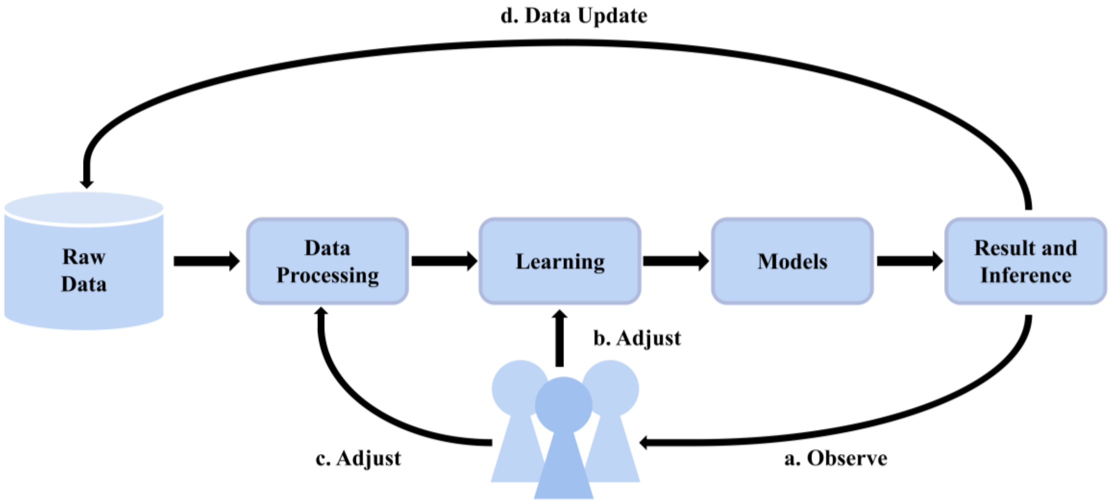

Figure 1. The concept of human-in-the-loop.

For the above difficulties, a new framework for introducing human knowledge into machines is proposed, namely human-in-the-loop (Wu et al., 2022) showed in Figure 1, and its effect has been demonstrated in a wide range of problem applications (Nunes et al., 2015). Based on the idea of human-in-the-loop (Zanzotto. F.M., 2019), this paper proposes a new two-stage feature processing framework for human-machine collaboration, which is successfully applied in movie box office prediction: the first stage is based on crawling from the public movie review platform. The movie box-office data of the film was screened by a regression decision tree algorithm to screen out the movie features that mainly affect the prediction results; in the second stage, an ANN network was built for box-office prediction. In particular, in the second stage of machine learning prediction, this paper introduces the prior knowledge of people in the prediction of movie box office revenue, manually classifies the features processed by the machine in the first stage, and divides multiple, incompatible feature sets to After designing the incompatible set edge pruning algorithm, the neural network is pruned (Reed. R., 1993), thereby reducing the noise caused by the direct coding between incompatible features (Kuhn et al., 2019), effectively improving the performance of machine learning methods in movies—prediction effect in the field of box office prediction.

RELATED WORK

Movie box office prediction has always been an essential issue in the economic field (Yu et al., 2012). Early movie box office predictions were mainly based on some artificially analyzed indicators, such as MPAA rating (Sharda et al., 2006) and the number of screens (Zhang et al., 2009). However, it is often difficult to explore the hidden laws in the data purely by using manual empirical analysis. The analysis results are not objective enough, leading to low prediction accuracy. To improve this shortcoming, some researchers want to perform box office prediction from publicly collected obtained data using a single data parsing method after analyzing the patterns, such as the multiple linear regression forecasting models (Subramaniyaswamy et al., 2017), Bayesian forecasting model (Ainslie et al., 2005), GBRT forecasting model (Han et al., 2018). In the analysis of Ainslie et al. (2005), it was found that the influence of specific actors and directors on the movie box office is very significant, which inspired this paper to design a unique process of screening important features before building a neural network for box office prediction. In addition, Liao et al. (2020) also integrated XGBoost, RF, LightGBM, KNN, and other machine learning methods to establish a stacking model for movie box office prediction.

In recent years, the vigorous development of computer technology has gradually transformed publicly collected movie feature datasets from structured to unstructured. Diverse data forms also prompt researchers to continuously explore new prediction methods, such as Duan et al. (2017) combining metadata and text data for feature-dependent box office prediction; Rajput et al. (2017) using Dual emotion Analysis Sentiment analysis performed on audience movie reviews; Han et al. (2017) propose a meta-heuristic method to analyze time-series data after movie release and discover different key features of box office growth through cluster analysis to improve prediction accuracy. Liu et al. (2019) proposed a hybrid algorithm that combines standard econometric methods with machine learning, achieving a good balance between short- and long-term prediction.

The development of deep learning methods has proven remarkable in several fields, such as Computer Vision (CV) and Natural Language Processing (NLP) (Pouyanfar et al., 2018, Srinivas, 2020 & Arif et al., 2021), while the earliest work to introduce them to movie box office prediction comes from Sharda et al. (2006), who used a multi-layer perceptron (MLP) for box-office revenues prediction, which consists of two hidden layers with a sigmoid activation function. Later, people complicated the network structure, and deeper networks were applied to movie box office prediction (Zhang et al., 2009 & Ghiassi et al., 2015), achieving better results. Since deep learning has a good ability to handle data with various modal information, the generalized regression neural network (IFOA-GRNN) model (Ru et al., 2019) combining the LSTM model (Ru et al., 2019) and virtual simulation calculation model (Lu et al., 2022) has also been successively applied to box office prediction. However, most publicly available movie datasets are still structured data that can be analyzed.

In addition, the movie box office prediction needs to focus on extracting some critical features. Combining the extracted features with a deep neural network is often better than building a deep neural network alone for prediction. Dai et al.(2019) propose an indoor location fingerprint recognition method combining a deep neural network (DNN) and an improved K-Nearest Neighbor (KNN) algorithm, which overcomes the limitation of the original KNN algorithm, which ignores the influence of neighboring points. Sun et al. (2019) calculated the confusion degree of emotion, then trained different DNNS for different emotion groups, extracted the bottleneck features for each SVM in the training decision tree, and realized the speech emotion classification. Phapatanaburi et al. (2016) apply the Gaussian mixture model (GMM) and deep neural network to recognize the speaker's accent in a reverberation environment, which is better than GMM and DNN alone due to their complementary effects. This paper; combines decision trees and ANNs to improve the box office prediction accuracy of large-scale structured movie public datasets.

As the research on machine learning progresses, more and more researchers recognize that introducing prior human knowledge to guide or constrain machine learning models can further improve the prediction. This is a Human-in-the-loop framework that has been introduced in Figure 1, i.e., introducing human a priori knowledge as an efficient way to process data into the learning model (Chen et al., 2020, Lin et al., 2020 & Hartmann et al., 2019), so that the model does not need to go through the data entirely for knowledge induction. Inspired by this, more and more researchers have begun to utilize a concept called “human-in-the-loop” to solve the problems of machines in various fields (Wu et al., 2022). In the field of movie forecasting, few papers are considering the introduction of the human-in-the-loop framework into box office forecasting. This paper is an earlier work that introduced the human-in-the-loop framework into movie box office prediction.

MODEL

The framework we propose is shown in Figure 1, which is mainly composed of two parts, including initial screening of features based on decision tree using data, removal of unimportant features, and artificial neural network pruning in the process of making theater box office predictions. The features are dealt with in detail. We also compare our framework with traditional statistical learning method and deep learning method in Figure 2. The specific features of the movie are shown in the Appendix I.

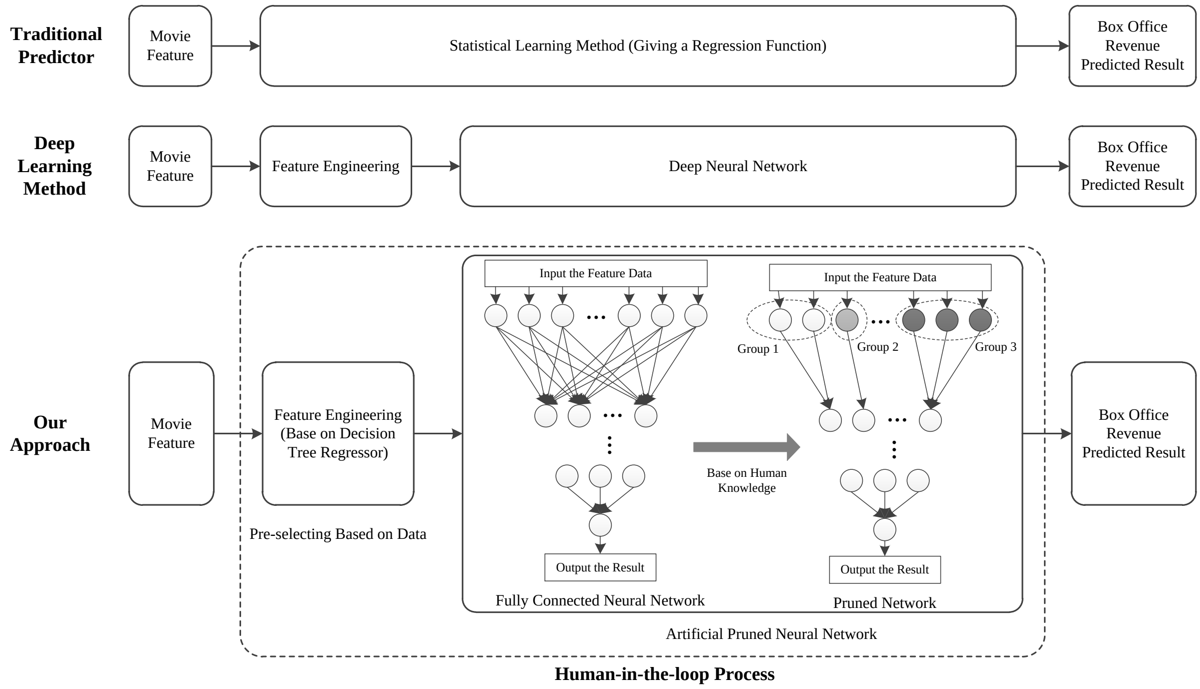
Traditional Predictor
Movie Feature
Statistical Learning Method (Giving a Regression Function)
Box Office Revenue Predicted Result
Deep Learning Method
Movie Feature
Feature Engineering
Deep Neural Network
Box Office Revenue Predicted Result
Our Approach
Movie Feature
Feature Engineering (Base on Decision Tree Regressor)
Pre-selecting Based on Data
Input the Feature Data
Output the Result
Fully Connected Neural Network
Base on Human Knowledge
Input the Feature Data
Group 1
Group 2
Group 3
Output the Result
Pruned Network
Artificial Pruned Neural Network
Box Office Revenue Predicted Result
Human-in-the-loop Process

Figure 2. The proposed method and its comparison with existed method.

3.1 Feature set preprocessing based on decision tree

The decision tree model (Kotsiantis S. B., 2013) is a general machine learning method based on the CART algorithm (Lewis R J, 2000), which mainly includes two parts: node structure and node split. Each node stores attributes that define the tree structure, including the left child and right child of the node, features, and threshold that are used for splitting the node. Denote the available movie feature set K. Denote the movie feature set after screening by the decision tree feature is K'. Denote the prediction result of the box office income of movie i at node m is y^i, $i \in m$. The decision tree calculates the loss function of CART under different feature k and threshold t_k, $k \in K$. After that, find out the value of feature and threshold that can minimize the loss function under the current node m, and obtain the feature of upper α through sorting as an effective screening feature.

$$J(k,t_k) = \frac{N_{left}}{N_m} MSE_{left} + \frac{N_{right}}{N_m} MSE_{right},$$
$$where \begin{cases} MSE_m = \sum_{i\in m} (\widehat{y_m} - y^i)^2 \\ \widehat{y_m} = \frac{1}{N_m} \sum_{i\in m} y^i \end{cases} \qquad (1)$$

N_m represents the number of samples of node m. N_{left} represents the left child node and N_{right} represents the right child node. Each node m will store the feature and threshold that minimize the mean square error (MSE), which is MSE_m.

Notes, in the process of solving the above tree model, the value of variable importance measures (VIM), which can measure the importance of movie features, can be solved together in the decision regression tree. The most commonly used expression of VIM is the Gini coefficient (Gastwirth, 1972) represented by GI. The Gini coefficient of each decision node is calculated as:

$$GI_m = 1 - \sum_{k=1}^{K} p_{m,k}^2 \qquad （2）$$

Among them, $p_{m,k}$ represents the proportion of movie feature k in the decision tree node m.

The importance of the movie feature X_k at node m can be expressed as：

$$VIM_{k,m}^{Gini} = N_m \times GI_m - N_{left} \times GI_{left} - N_{right} \times GI_{right} \qquad （3）$$

Then the importance of X_k in decision regression tree can be accumulated as：

$$VIM_k = \sum_{m \in M} VIM_m \qquad （4）$$

The hyperparameter that defines the boundaries of expression feature filtering is α, then the feature screening process aimed at determining the set can be expressed by equation (5)

$$\underset{K' \subseteq K}{arg} \ (VIM_k \geq \alpha, k \in K') \qquad （5）$$

3.2 Film box office prediction model based on ANN

Neural networks are currently widely used in the modeling of complex input-output relationships for classification and prediction problems. This paper constructs the box office revenue prediction problem as an ANN neural network prediction model. After the tree model preprocessing, the input layer size is equal to $|K'|$, including ten hidden layers, the width of each hidden layer is the same, and the output layer size is 1, that is, the predicted value of the theater box office.

Defined z_i^l as the i-th node in layer I and $w_{i,j}^l$ as the weight between the i-th node in layer I and the j-th node in layer $I+1$, b_i^l is the bias term of the i-th node in layer l. Then, for any neuron i, the value iteration in forward pass process can be represented as equation (6)：

$$z_j^{l+1} = \sigma\left(\sum_{i=1}^{I} w_{i,j}^l z_i^l + b_j^{l+1}\right) \qquad （6）$$

Among them, σ is activation function.

Define the neuron output of the network output layer as y. For any sample feature x_k mapped by ANN to y as $y = net(x_k)$, the corresponding loss function is defined as:

$$L = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_i - net(x_k))^2} \qquad (7)$$

By using gradient descent (Ruder, 2016), the model parameters of the box office revenue prediction ANN network will converge, and then the specific box office revenue value of a film can be predicted.

3.3 Artificial Network Pruning

Based on human prior knowledge on box office revenue predictions, we prune the edges of the ANN during the forward pass. We define Q incompatible sets H , $H = \{H_1, H_2, ..., H_Q\}$ For the first layer, the nodes in the hidden layer are $x_i^{L1}, i \in H_q$. After pruning, its forward pass process can be expressed as：

$$x_i^{L1} = \sigma(\sum_{h \in H_q} w_{i,h} + b_{L1}) \qquad (8)$$

The specific rules are as follows:
We are considering that when the film is released at different times and genres, the film's audience will also be different. For example, the student group strongly desires to consume movies during the summer vacation, so the box office of movies with themes popular with the youth group, such as comedy and fantasy, will be satisfying during the summer vacation season. Thus, we believe there is a link between film genre and slotting.

This paper categorizes comedy, fantasy, romance, sci-fi, original, based on characters, based on legend, based on actual events as the features of the film genre and categorizes May Day, New Year's Day, National Day, Christmas, summer July, and Qingming as features of the movie schedule. Because the number of competitors and similar competitors during the schedule is related to the specific schedule, consider the above three and delete the propagation paths between the three types of topics and other features to the next hidden layer.
Additionally, we combine the features of the same subject.

1. Director Features. Through observation, we found that the four characteristics of "total box office of director's films, the box office of director's films, number of director's works, and number of director's nominations" are all subject to film directors, and the intrinsic connection between them is significant. For example, the box office of a director's single film strongly impacts the number of director award nominations. Therefore, we output these four features to three specific hidden layer neurons to obtain results characterizing the director's influence on the movie box office.

2. Actor-based features. We calculated the number of historical works of actors, the average number of works, the total box office of actor’s works, the average box office of works, the box office of single history works of actors (the sum of the single works of an actor/the

number of actors), the total box office/number of works, the total number of nominations, the average number of nominations, the total number of awards, and the average number of awards are classified as actor-based features.

3. Film production company features. We classify the number of production companies, the number of historical works, the average number of historical works, the historical box office, the number of historical box office/historical works, the average historical box office, and the average historical single box office as the characteristics of the production company.

4. Film distribution company features. We categorize the number of historical titles, average historical titles, historical total box office, historical box office/historical titles, average historical total box office, and average historical single box office of the distribution companies as the main characteristics.

5. Movie trailer traffic features. The average value of Weibo's trailer playback, forwarding, likes, comments, poster click-through rate, and the comprehensive score is classified as features, with trailer traffic as the main feature.

6. Movie-related topics traffic features. The average value of Tik Tok collection, the average value of Weibo picture comments, the average value of likes, the average value of Tik Tok popularity, the average value of account influence, the average value of video discussion, the average value of topic popularity, the average value of search index, and the average value of Weibo total score and the average size of the population covered by Weibo and the average Weibo discussion degree are classified as the features of the traffic of movie-related topics.

7. Market features. The total number of competitors in the same period, competitors of the same type in the same period, the length of the historical schedule, and the box office of the historical schedule are classified as the features of the market.

Considering that directors, actors, production companies, distribution companies, and social platform traffic-related data subjects are different and have a low degree of correlation, the output paths of different types of input layer neurons to some neurons in the first hidden layer are selectively deleted.

Due to the different nature of trailers and posters from other related pictures on Weibo and Tik Tok, the former emphasizes the overall effect and plot of the film and is a more official way of publicity, while the latter is more entertainment and discussion, including publicity, and secondary creation can more intuitively show the willingness and enthusiasm of the masses to consume so that we will process the two separately.

To summarize, the detail of the proposed method is shown in Table 1 as follow:

Table 1. Pseudocode.

```
Pseudocode of the proposed approach
Input: Movie office box data and the set of movie feature $K$ .
Construct the decision tree function $f_{tree}$ based on equation (1) using the input data and the feature set $K$ .
    for $k$: $=1$ to $|K|$
        Calculate $VIM_k$ for each feature $k$ based on equation (2) to equation (4).
    Construct new feature set $K'$ based on VIM and equation (5).
    Construct fully-connected neural work $f_{net}$ based on equation (6).
    Prune the network $f_{net}$ based on equation (8) and the rule purposed in 3.4
    Train the network $f_{net}$ based on equation (7).
return $f_{net}$ ,$f_{tree}$ ,$K'$
```

Experience

Data Description: Using crawler technology, we constructed a dataset of 7098 movie online data crawled on Douban, Weibo, Douyin, Lighthouse, and other platforms. After data cleaning, each piece of data includes a total of 58 features except movie names (including three discrete features and 55 continuous features), labeled as movie cumulative box office.

Metrics: After normalizing and post-processing the data, we use Mean Absolute Error (MAE) as the metric for regression prediction, and we introduce regression decision tree, linear regression (Mestyán et al., 2013), and support vector regression algorithm (Liu et al., 2016) as baselines to compare with our algorithm.

Experimental Settings. The proposed method is trained by using the Adadelta optimizer (Zeiler et al., 2012) to minimize the Mean Square Error (MSE). In addition, total training epochs are set to 5000. We run the experiments on a compute node with an Intel Xeon Silver 4210R (40) @ 3.200GHz and 128 GB of RAM. We use a GPU node with a NVIDIA GeForce RTX 3090 Ti Accelerator Card to train deep learning models with the support of the Pytorch library (Paszke et al., 2019).

4.1 Feature Screening and Analysis

This paper extracts the total box office of actors and directors, the box office of a single work, the number of awards and nominations, the number of trailers, the number of reposts, the number of likes and comments, and the flow of movie-related graphics, film reviews, and second-generation works. Data and other 58 features are quantified, three discrete character features of schedule, type, and adaptation type are subjected to one-hot encoding processing, and a total of 103 features are finally obtained. The analysis results of the feature scores through the decision tree are as Figure 3 shows:

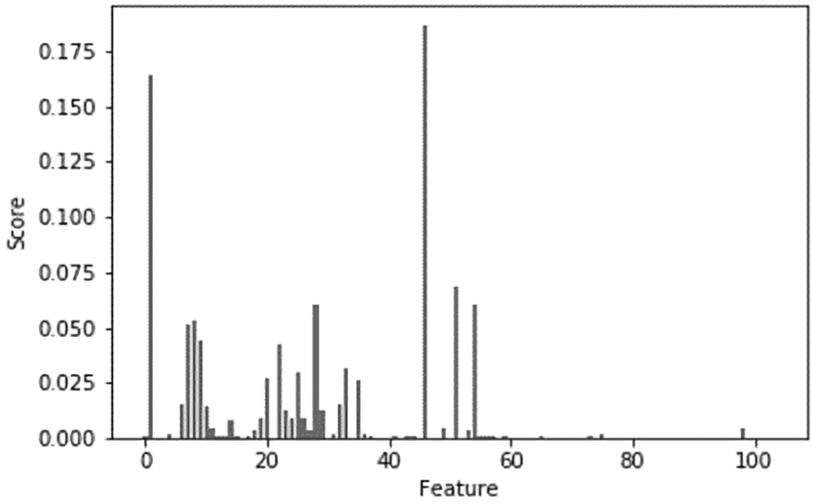

Figure 3. Feature screening based on decision tree.

This paper extracts the total box office of actors and directors, the box office of a single work, the number of awards and nominations, the number of trailers, the number of reposts, the number of likes and comments, and the flow of movie-related graphics, film reviews, and second-generation works. Data and other 58 features are quantified, three discrete character features of schedule, type, and adaptation type are subjected to one-hot encoding processing, and a total of 103 features are finally obtained. The analysis results of the feature scores through the decision tree are shown in Figure 4.

We analyzed the features of the main effects. Among the 71 features, the ten most influential typical features are screened out; their numbers and importance are shown in the Figure 4. The abscissa represents the feature number in the processing process, and the ordinate represents the feature importance coefficient.

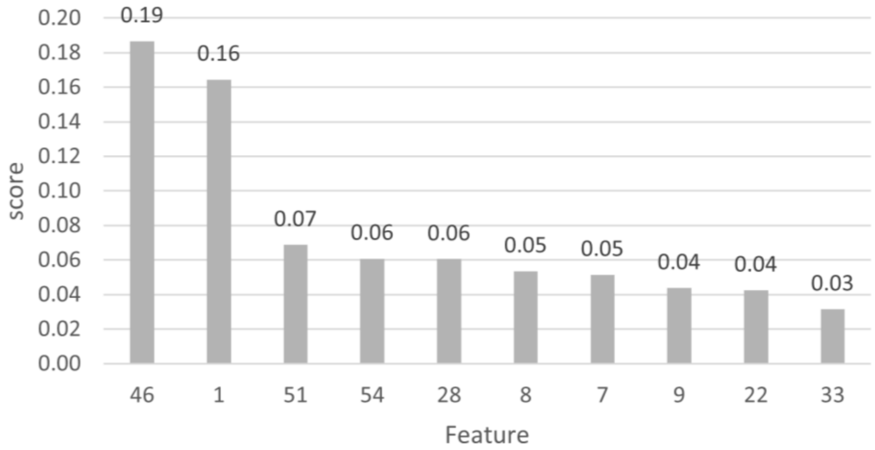

Figure 4. The 10 most influential typical features.

Among them, the No. 46 feature is the Weibo search index, and its importance is 0.19, which has the most significant impact on the box office. The No. 28 and 33 factors, ranked fifth and tenth in importance, are also related to the Weibo platform. It shows that the search volume of social media such as Weibo and the flow of movie-related topics reflected by the recognition degree has a huge impact on the movie box office.

The second most important No.1 feature is the box office of the director's single work, and the sixth, seventh, and eighth most important features are actors-related factors, which are the total box office of the actor's historical works, historical works The average box office and the box office of the actor's history of a single work (the sum of the single work per actor/number of actors), which means that the film production participants have a huge influence on the film box office. On the one hand, the past performance of the participants shows that The level of its production and participation, on the other hand, also reflects its mass base, star effect, brand effect, etc., all of which provide certain support for the prediction of the movie box office.

The third and fourth most essential features are whether the film is a sci-fi type film and whether it is an original film, which means that in recent years, audiences have a strong willingness to watch sci-fi series films and prefer original works.

The ninth most important is the number of historical works of the production company, which shows that as the investor and beneficiary of the film, the production company also has a significant influence on the film box office. At the same time, the production company has a specific brand effect, and its social influence also affects the box office to a certain extent.

4.2 Performance and Comparison

We built a deep neural network with an input layer size of 71, corresponding to the number of features filtered through the decision tree model, a total of ten hidden layers, the size of the hidden layer is uniformly set to 40, and the output layer size is 1 to output the box office revenue of the theater, we use the selected features as the input of the pruned ANN. As shown in the Figure 5 as below, the model finally converges during the training process, and its accuracy on the test set reaches 79.93%.

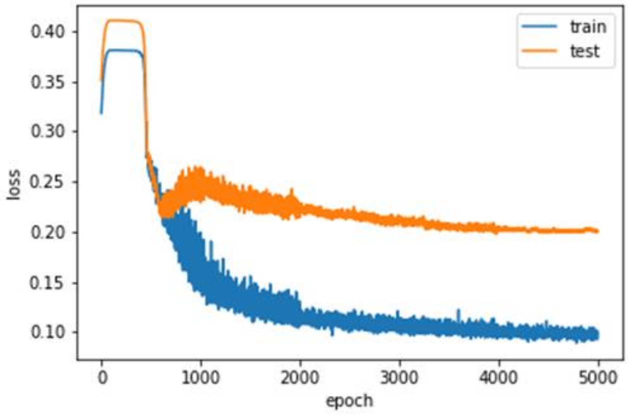

Figure 5. ANN training process.

The comparison results of our method and the baseline are shown in the Table 2. It can be seen that the linear regression method has the weakest performance in the box office prediction problem, and the accuracy rate is only 64.45%, while the regression decision tree and support vector regression methods have relatively higher accuracy. accuracy (71.81% and 72.68%), and the proposed method can achieve the highest accuracy of 79.93%.

Table 2. Comparison with baseline method.

	Our method	Linear regression	Decision tree	SVM
Accuracy (%)	79.93	64.45	71.81	72.68

4.3 Ablation Study

To demonstrate the effectiveness of the human in the loop method, we experimentally compare our method with the ANN method without manual pruning. That is, both perform the same feature screening based on decision regression trees and then filter. The output features are used as the input of the next ANN. However, the connection of the neurons in the first two layers of the former ANN after manual pruning is relatively sparse, while the latter is fully connected. The comparison results are shown in the Table 3:

Table 3. Comparison of the prediction results of the model with or without the introduction of the human-in-the-loop framework.

	ANN with artificial pruning	ANN without artificial pruning
Accuracy(%)	79.93	72.35

It can be seen that the result of our method after adding human prior knowledge to prune the network is obviously higher than the result obtained without pruning, which proves the effectiveness of our method. Comparing the results of the fully connected ANN combined with the decision tree model with the results of using the decision tree method alone in table1 (72.35% vs. 71.81%), we can also confirm that the combined framework of the tree model combined with the deep neural network is better than using the tree alone effect of the model.

CONCLUSION

We propose a prediction algorithm of "feature screening-neural network construction-artificial pruning" that draws on the design idea of "human-in-the-loop" for human-machine collaboration. It combines traditional machine learning and deep learning theories and is used in the field of movie box office revenue prediction for the first time. The algorithm introduces prior human knowledge based on the decision tree algorithm that uses traditional machine learning for feature processing based on data, artificially prunes the constructed deep learning ANN network, and then trains the pruned ANN network to obtain a predictive parametric model. The experimental results show that this paper's algorithm improved the algorithm significantly in prediction accuracy compared with traditional algorithms such as linear regression, decision trees, and support vector machines commonly used in the industry. This proves the effectiveness of the method proposed in this paper. That is, the human-in-the-loop is realized by combining the advantages of humans and machines. As the first article to introduce the idea of human-machine collaboration into the design process of movie box office revenue prediction algorithms, this work also reveals that AI algorithm design based on human-machine collaboration has broad application prospects in various fields. The follow-up work will explore the human-in-the-loop prediction method under unstructured data sets, and we will also try to practice it in the movie box office revenue management. Otherwise, we will continue to explore the particular problem structure of prediction and classification tasks in other application backgrounds, analyze the unique role played by humans, and design higher-precision human-machine collaborative machine learning algorithms accordingly.

ACKNOWLEDGEMENT

The total authors wish to thank the three undergraduates form Zhejiang University, who are Siqi Ye, Shengdong Wu and Liruoyun Pang, on the manuscript translation. We also wish to thank the vice-headmaster Weibin Weng, Dr. Yu Zhou and Dr. Jingwei Yao on the initiation of the research training program for high school students. The research database is authorized to extract from the International Research Center for Data Analytics and Management, Zhejiang University. We appreciate the support from National Natural Science Foundation of China (Project Number: 72192823).

REFERENCES

Kiran, B.R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A.A., Yogamani, S. & Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems.1-18.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T. & He, L. 2022. A survey of human-in-the-loop for machine learning. Future Generation Computer Systems. 135 364-381.

Zanzotto, F.M. 2019. Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research.64 243-252.

Reed, R. 1993. Pruning algorithms-a survey. IEEE transactions on Neural Networks.4 (5) 740-747.

Wang, J., Xu, C., Huangfu, Y., Li, R., Ge, Y. & Wang, J. 2019. Deep reinforcement learning for scheduling in cellular networks. In 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP).

Nunes, D.S., Zhang, P. & Silva, J.S. 2015. A survey on human-in-the-loop applications towards an internet of all. IEEE Communications Surveys & Tutorials.17 (2) 944-965.

Kuhn, M. & Johnson, K. 2019. Feature engineering and selection: A practical approach for predictive models. CRC Press.

Kotsiantis, S.B. 2013. Decision trees: a recent overview. Artificial Intelligence Review.39 (4)

261-283.

Lewis, R.J. 2000. An introduction to classification and regression tree (CART) analysis. Annual meeting of the society for academic emergency medicine in San Francisco. California, Citeseer.

Gastwirth, J.L. 1972. The estimation of the Lorenz curve and Gini index. The review of economics and statistics. 306-316.

Zeiler, M.D. 2012. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, F., Bai, J. & Chintala, S. 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems.

Mestyán, M., Yasseri, T. & Kertész, J. 2013. Early prediction of movie box office success based on Wikipedia activity big data. PloS one.8 (8) e71226.

Liu, T., Ding, X., Chen, Y., Chen, H. & Guo, M. 2016. Predicting movie box-office revenues by exploiting large-scale social media content. Multimedia Tools and Applications.75 (3) 1509-1528.

Han, J.M., Hong, M., Jung, J.J. & Camacho, D. 2017. Metaheuristic approach on temporal pattern matching for box office prediction. Proceedings of the International Conference on Electronic Commerce.

Duan, J., Ding, X. & Liu, T. 2017. A Gaussian copula regression model for movie box-office revenues prediction. Science China Information Sciences.60 (9) 1-14.

Rajput, P., Sapkal, P. & Sinha, S. 2017. Box office revenue prediction using dual sentiment analysis. International Journal of Machine Learning and Computing.7 (4) 72-75.

Han, T., Yuan, B., Chen, Y, Zhao, N. & Duan, D. 2018. An effective box office prediction model based on GBRT. Computer Application Research.35 (2) 410-416.

Liu, Y. & Xie, T. 2019. Machine learning versus econometrics: prediction of box office. Applied Economics Letters.26 (2) 124-130.

Liao, Y., Peng, Y., Shi, S., Shi, V. & Yu, X. 2020. Early box office prediction in China's film market based on a stacking fusion model. Annals of Operations Research.1-18.

Zheng, J. & Zhou, S. 2014. Modeling of movie box office prediction based on neural network. Computer Applications.34 (3) 742-748.

Ru, Y., Li, B., Chai, J. & Liu, J. 2019. A daily box office prediction model with LSTM. 2019 International Conference on Artificial Intelligence and Advanced Manufacturing.

Lu, W., Zhang, X. & Zhan, X. 2022. Movie box office prediction based on IFOA-GRNN. Discrete Dynamics in Nature and Society. 2022.

Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. 2019. Speech emotion recognition based on dnn-decision tree svm model. Speech Communication. 115.

Dai, P., Yang, Y., Wang, M. & Yan, R. 2019. Combination of DNN and improved KNN for indoor location fingerprinting. Wireless Communications and Mobile Computing. 2019.

Phapatanaburi, K., Wang, L., Sakagami, R., Zhang, Z., Li, X. & Iwahashi, M. 2016. Distant-talking accent recognition by combining GMM and DNN. Multimedia tools and applications.75 (9) 5109-5124.

Yu, S. & Kak, S. 2012. A survey of prediction using social media. arXiv preprint arXiv:1203.1647.

Sharda, R. & Delen, D. 2006. Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications.30 (2) 243-254.

Zhang, W. & Skiena, S. 2009. Improving movie gross prediction through news analysis. International Joint Conference on Web Intelligence and Intelligent Agent Technology.

Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V. & Logesh, R. 2017. Predicting movie box office success using multiple regression and SVM. 2017 international conference on intelligent sustainable systems (ICISS).

Liao, Y., Peng, Y., Shi, S., Shi, V. & Yu, X. 2020. Early box office prediction in China's film market based on a stacking fusion model. Annals of Operations Research.1-18.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M., Chen, S. & Iyengar, S.S. 2018. A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR).51 (5) 1-36.

Srinivas, V. 2020. LFBNN: Robust and Hybrid Training Algorithm to Neural Network for Hybrid Features-Enabled Speaker Recognition System. Journal of Engineering Research.8 (2).

Zhang, L., Luo, J. & Yang, S. 2009. Forecasting box office revenue of movies with BP neural network. Expert Systems with Applications.36 (3) 6580-6587.

Ghiassi, M., Lio, D. & Moon, B.2015. Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Systems with Applications.42 (6) 3176-3193.

Chen, S., Leng, Y. & Labi, S. 2020. A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information. Computer-Aided Civil and Infrastructure Engineering.35 (4) 305-321.

Lin, Y., Pintea, S.L. & Gemert, J.C.V. 2020. Deep hough-transform line priors. In European Conference on Computer Vision. Springer, Cham.

Hartmann, G., Shiller, Z. & Azaria, A. 2019. Deep reinforcement learning for time optimal velocity control using prior knowledge. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI).

Arif, S. & Wang, J. 2021. Bidirectional LSTM with saliency-aware 3D-CNN features for human action recognition. Journal of Engineering Research.9 (3A).

Ruder, S. 2016. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Ainslie, A., Drèze, X. & Zufryden, F. 2005. Modeling movie life cycles and market share. Marketing science.24 (3) 508-517.

Appendix I. A Sample of Specific Movie Data in the Research Database

Movie title	The Battle at Lake Changjin
Movie type	War
Type of adaptation	True History
Box office revenue of all the director's former works	4748743590.0
Box office revenue of the director's each former work	395728632.5
Number of the director's former works	12
Nominations number of the director	12
Awards number of the director	6
Number of actors	3
Number of the actors' former works	26
Average of each actors' former works	8.7
Box office revenue of all the actors' former works	20140407480.0
Box office revenue of each actor's former work	6713469160.0
Average box office revenue of all the actor's former work	785869635.0
Average box office revenue of each actor's former work	774631056.9
Nominations number of the actors	14
Nominations number of each actor	4.7
Awards number of the actors	9
Awards number of the actors	3
Number of movie productors	3
Number of all the movie productors' former works	58
Number of the movie productors' former works	19.3
Box office revenue of all the movie productors' former works	14220704206.0
Box office revenue of all the movie productors' former works	245184555.3
Average box office revenue of all the movie productors	4740234735.0
Average box office revenue of each movie productors' former work	507387273.7
Number of movie distributors	3
Number of all the movie distributors' former works	917
Number of the movie distributors' former works	305.6666667
Box office revenue of all the movie distributors' former works	147480000000.0
Box office revenue of all movie distributors' former works	160829037.8
Average box office revenue of all movie distributors	49160075895
Average box office revenue of each movie distributors' former work	244733650.7
Trailer views on Weibo	9123333.3
Amount of reposting of each Trailer on Weibo	32112.5
Amount of likes of each Trailer on Weibo	59532.7
Amount of comments of each Trailer on Weibo	9275.3
Amount of reposting of relevant pictures on Weibo	8116.3
Amount of likes of relevant pictures on Weibo	2154.8
Amount of comments of relevant pictures on Weibo	13087.4
Trailer views on Tik Tok	461344.2
Amount of reposting of each Tik Tok	13926.3
Amount of likes of each Tik Tok	4910.1

Amount of reposting of each Tik Tok	11396.1
Poster views	0.112
Heat on Tik Tok	16649791.7
Influence of account on Tik Tok	1887687.5
Heat of the videos on weibo	1471625.0
Heat of the topics on weibo	976708.3
Heat of search on weibo	13505739.1
The comprehensive popularity of Weibo	89.0
Covered population on weibo	32.5
Heat of Weibo discussion	29.4
Trailer views on Weibo	28.2
Window	National holidays
Competitors during the same window	8
Competitors of same type during the same window	2
Length of window	6.3
Box office revenue of all former works during the window	2509568571
Box office revenue	5741362162