Detection of anomalous ticket purchasing behavior for concerts based on machine learning

Zeng Lingyang[1,2], Peng Lilin[1]

[1]Brunel London School, North China University of Technology, Beijing, China

[2]Corresponding author: zly1357911@163.com

Abstract. In recent years, there has been a continuous increase in the demand for entertainment activities. With the changing policies related to epidemic prevention and control, consumers' ticket demands for various entertainment events, particularly concerts, have seen a significant rise. This has led to the emergence of anomalous ticket purchasing behaviors, including ticket scalping software, bulk purchasing, and fraudulent transactions. Such behaviors not only infringe upon the rights of legitimate audiences but also harm the interests of concert organizers. Therefore, the detection and prevention of anomalous ticket purchasing behaviors for concerts have become essential. This paper focuses on the use of automated systems, robots, or malicious software for ticket purchases, limiting consumers' participation in abnormal ticket-buying activities. Combining previous research on railway ticketing systems and real-life experiences, this study selects ticket-purchasing frequency, speed, quantity, seat selection, payment method, IP address changes, and historical ticket-purchasing records as feature values. Different machine learning models are chosen for different feature values to conceptualize the construction of an anomalous ticket purchasing system for concerts. This paper emphasizes the importance of detecting anomalous concert ticket purchasing behavior, provides different machine learning models for various feature values, and lays the foundation for building an overall anomaly detection system. This research serves as a crucial reference for the security and user experience of ticketing systems, aiming to continually improve and upgrade anomaly detection systems to adapt to evolving challenges. The hope is to enhance the ticket-buying experience for future concerts and large-scale events.

Keywords: Machine Learning, Anomaly Detection, Abnormal Concert Ticket Purchase Behavior

1. Introduction

With the continuous development of the music industry and the popularity of concerts, coupled with the gradual easing of the pandemic this year, there has been an increasing demand for entertainment activities. In just the first half of 2023, more than 840 individual concerts can be found on ticket platforms. However, due to the limited availability of concert tickets, the ticket purchasing process is often accompanied by intense competition, providing opportunities for the occurrence of abnormal ticket purchasing behavior, such as ticket-scalping software, bulk purchasing, and fraudulent transactions. This not only makes it difficult for legitimate audiences to obtain tickets but also causes losses to the ticketing systems and organizers of the concerts. Therefore, the detection and prevention of abnormal ticket purchasing behavior for concerts have become crucial. The abnormal ticket purchasing behavior studied in this paper refers to the use of automated systems, robots, or malicious software to snatch tickets, thereby restricting the participation of regular attendees.

Abnormal ticket purchasing behavior at concerts can have negative impacts on both the concert industry and the audience. For concert organizers, such behavior may lead to unfair ticket distribution and even result in the collapse of ticketing systems, potentially affecting the entire ticketing market in the long run. For the audience, abnormal ticket purchasing behavior may affect the normal use, causing a decline in user satisfaction. Hence, in-depth research and resolution of abnormal ticket purchasing behavior for concerts hold significant practical significance. Past studies have explored the use of machine learning techniques in various fields to detect abnormal behavior, such as credit card fraud detection and network intrusion detection. These studies provide valuable experiences that can be applied to the field of concert ticket sales. In this context, this paper aims to propose a machine learning-based approach for automatically identifying and preventing abnormal ticket purchasing behavior at concerts, focusing on model selection and construction due to user information privacy concerns. Building upon existing models for identifying user abnormal behavior and railway ticket-snatching, this paper transforms these models to be applicable to the concert ticket-snatching environment. The paper seeks to delve into methods for model selection and construction, providing a robust approach to the identification and prevention of abnormal ticket purchasing behavior by recognizing feature values. This, in turn, is expected to promote the sustainable development of the concert industry, uphold fairness in the concert ticket market, and advance the application of machine learning in the security domain.

2. Literature Review

2.1. Definitions Related to Anomaly Detection

Anomaly detection, also known as outlier detection, is the process of identifying objects that deviate from the majority of objects, essentially identifying outliers. Anomalies are sometimes referred to as deviations.

An anomaly, or outlier, is an observed data point that is far from other data points and is suspected to have originated from a different mechanism. Anomalies can be categorized into three main types: point anomalies, conditional anomalies (or contextual anomalies), and collective anomalies (or group anomalies).

Anomaly detection methods are generally classified into three types: unsupervised, supervised, and semi-supervised. It is a process of identifying elements that do not conform to expected patterns or other items in a dataset. Anomaly detection finds wide applications in various domains, including intrusion detection, fraud detection, fault diagnosis, system health monitoring, and anomaly detection in timeseries data.

2.2. Anomaly Detection Methods

2.2.1. Unsupervised Anomaly Detection

The core idea of unsupervised anomaly detection techniques is that anomalous situations are rare compared to normal situations, and they exhibit significant differences in aspects such as data distances, distribution density, and deviation.

Unsupervised anomaly detection algorithms can detect outliers in data based solely on intrinsic properties such as distance and density. Autoencoders serve as the core of all unsupervised deep anomaly detection models.

In the field of unsupervised anomaly detection, data sets can be broadly categorized into two types: semantic-level anomaly detection and region-level defect detection. Semantic-level anomaly detection is essentially the task of requiring the model to distinguish samples with semantic-level anomalies. On the other hand, region-level defect detection is primarily applied in the industrial quality inspection domain, such as detecting anomalies in industrial components. From the perspective of the entire image, there may be no significant changes (still belonging to the same category), but defects may exist in a specific local region. Compared to semantic-level anomaly detection, tasks of this nature are more challenging. However, they hold practical significance in real-world applications, particularly in industrial contexts, and have therefore garnered widespread attention from researchers in the past two years.

Wei et al. proposed an anomaly detection method in their paper “Unsupervised anomaly detection by densely contrastive learning for time series data.” This method involves dense contrastive learning in the latent space for entire time series and sub-sequences at different timestamps. It effectively captures local features of sub-sequences using the locality of convolutional neural networks (CNN) and position embeddings [1]. Additionally, attention mechanisms are employed to extract global features from the entire time series. The model is trained using instance-level contrastive learning loss and distribution-level alignment loss. A reconstruction loss for extracting global features is also introduced to prevent potential information loss. The effectiveness of the proposed technique is validated through anomaly detection experiments on publicly available time series datasets.

Jin et al. introduced a lightweight unsupervised anomaly detection scheme for multivariate time series data in their paper “LUAD: A lightweight unsupervised anomaly detection scheme for multivariate time series data.” LUAD comprises a detection model and a diagnostic model [2]. The detection model combines the encoder-decoder scheme of time convolutional networks (TCN) and variational autoencoders (VAE) to learn the normal patterns of input data, deconstructing and reconstructing multivariate time series data. The diagnostic model enhances LUAD’s overall detection accuracy and provides reasonable explanations for anomalies.

2.2.2. Supervised Anomaly Detection

Supervised anomaly detection requires a dataset that has already been labeled as “normal” and “anomalous,” involving the training of a classifier (with a key distinction from many other statistical classification problems being the inherent imbalance in anomaly detection).

In his paper on the research of anomaly detection in process data based on supervised learning, Kun Kun Wang focuses on introducing the principles and applications of ensemble learning frameworks in machine learning algorithms. The paper also covers several methods of data feature dimensionality reduction, particularly highlighting principal component analysis (PCA) and stacked autoencoder methods. In the paper, a feature selection method that combines PCA with ensemble decision trees is chosen to rank the importance of features in industrial data. The study explores an anomaly detection method based on an ensemble framework. Data collected from industrial sites undergo feature engineering, employing a combination of PCA and ensemble decision trees [3]. The paper utilizes random forest (RF), Adaboost, Xgboost, and SVM as base learners, with logistic regression (LR) as the meta-learner to construct a Stacking ensemble method for anomaly detection. The model is trained using TE data, and the experimental results of the ensemble model are compared with those of the base models, effectively improving the accuracy of anomaly detection in process data and reducing the false positive rate.

In the paper “Anomaly detection in NetFlow network traffic using supervised machine learning algorithms,” Igor et al. tested several classification algorithms (stochastic gradient descent (SGD), support vector machine (SVM), K-nearest neighbors (K-NN), Gaussian naive Bayes (GNB), decision tree (DT), random forest (RF), AdaBoost (AB)) on the UNSW-NB15 public dataset [4]. Different encoding methods and the ratio of training data to testing data led to the optimal parameter classifiers. The paper compares various anomaly detection algorithms, selecting the most suitable one for anomaly detection in NetFlow data streams. F2-score and AUC metrics were applied, using label encoding (LE). The paper optimizes the machine learning process for NetFlow data streams, considering different ratios of testing data and applying various encoding methods and feature reduction techniques.

2.2.3. Semi-Supervised Anomaly Detection

Semi-supervised anomaly detection involves creating a model that represents normal behavior based on a given training dataset labeled as "normal." The model is then used to assess the likelihood of test instances generated by the learning model.

A typical approach in semi-supervised anomaly detection is to build a model on a training dataset of normal time series data and use this model to identify anomalies in the test data. Semi-supervised anomaly detection learns the discriminative boundary of normal data, classifying data not belonging to the normal class as anomalies. Due to the absence of labeled anomaly sequences during training, the application of semi-supervised anomaly detection methods is relatively more widespread.

In the paper on anomaly detection in student consumption data based on semi-supervised learning, Song et al addressed the issue of traditional anomaly detection methods being ineffective in extracting temporal features from student consumption data. They proposed a semi-supervised learning-based method for detecting anomalies in student consumption data [5]. Firstly, they improved the autoencoder with gated recurrent units to enhance the accuracy of reconstructing consumption data. Then, they employed the Mahalanobis distance to calculate the reconstruction error, computed the Fβ-score to determine the error threshold, and conducted anomaly detection. Finally, they applied the proposed method to conduct anomaly detection experiments on the consumption data of students in a certain university..

2.3. Related Research on Anomaly Detection

In the paper "Research and Application of Abnormal Behavior Classification Technology in Railway Internet Ticket Sales," Zhou et al. collect user ticket purchase logs, transaction records, IP address tracking, and other data. They select features such as ticket purchase frequency, purchase time distribution, purchase location, and payment methods for model training and anomaly behavior classification using machine learning and data mining techniques. The model's performance metrics are used to evaluate its effectiveness in anomaly behavior detection [6].

Li et al., in their paper "Research and Implementation of Intelligent Identification of Abnormal Users in Railway Internet Ticket Sales Based on Exponential Weight Algorithm," gather user ticket purchase behavior logs, transaction records, and purchase time distribution data for model construction. They explore the application of the exponential weight algorithm in identifying abnormal users [7]. This algorithm is commonly used for modeling user behavior, balancing the importance of actions to help distinguish between normal and abnormal users.

In the paper "Analysis of Dishonorable Behavior On Railway Online Ticketing System Based on k-Means and FP-Growth," Yang et al. propose a model for detecting abnormal ticket purchase behavior in a railway online ticketing system using traditional K-Means and FP-Growth algorithms [8]. The random forest algorithm based on Spark MLlib is employed for preliminary feature selection, identifying features closely related to user ticket purchase behavior. Subsequently, K-Means clustering analysis is applied to each selected feature, assigning each user feature to a specific type. Finally, FP-Growth is used to determine multiple high-precision feature combinations for identifying abnormal behavior.

Peng et al., in their paper "Research on User Anomaly Detection Based on SVM under Time Sequencing," suggest using SVM for preliminary anomaly information detection, followed by Bayesian transformation for time sequencing [9]. By integrating machine learning and text analysis, the study aims to enhance the understanding of the commonalities and characteristics of anomaly data, facilitating the correlation analysis of multidimensional relationships within integrated anomaly data. This approach simplifies and improves post-monitoring maintenance work and smooths the network environment in the later stages.

These relevant studies provide various reference models for the detection and classification of abnormal online ticketing behavior in this research. They offer preliminary ideas and insights for feature selection and model construction, suggesting a method to allocate weights to the importance of each feature, thereby aiding in the detection of abnormal ticketing behavior. Additionally, the studies provide detailed insights into the Spark MLlib-based random forest algorithm and SVM classification algorithm.

They offer practical methods for the initial screening of user features and the identification of relevant features in this research. Besides, detailed instructions for the usage of the K-Means algorithm and FP-Growth algorithm are provided, offering an approach for clustering analysis of data and determining feature combinations in this research. What's more, the studies also introduce the Bayesian transformation, providing a method for time sequencing in the detection of the feature value related to ticketing speed. Simultaneously, leveraging the integration of machine learning and text analysis methods, these studies offer references for analyzing the commonalities, characteristics, and relationships among abnormal data in this research.

3. Model Feature Selection

Anomalies in concert ticket purchasing behavior include the rapid acquisition of a large number of tickets within a short time frame, frequent changes in payment methods within a short time, frequent ticket purchasing behavior within a short time, frequent changes in IP addresses during ticket purchases, ordering quantities exceeding the limit using batch identical or fictitious payment accounts, delivery addresses, recipients, phone numbers, ordering quantities exceeding the limit using the same ID, or excessively short usage time at each ticket purchase stage compared to other users. Therefore, this study selects the following model features.

Table 1. Model Feature Selection

Feature	Explanation
Ticket Purchase Frequency	Indicates whether the user engages in frequent ticket purchases within a short period.
Ticket Purchase Speed	Examines if the user's usage time at each ticket purchase stage is excessively short compared to other users.
Ticket Purchase Quantity	Evaluates whether the user is involved in a large quantity of ticket purchases.
Ticket Seat Selection	Checks if the user repeatedly purchases the same seat or consistently chooses specific seats.
Payment Method	Identifies the presence of unusual payment methods or frequent changes in payment methods.
IP Address Changes	Determines if the user frequently changes IP addresses or uses multiple IP addresses for ticket purchases.
Historical Ticket Purchase Records	Assesses whether the user has a history of abnormal behavior in ticket purchases.

4. Model Selection and Construction

Due to the sensitive nature of concert user data involving a significant amount of user privacy, it is not possible to obtain real user purchasing records for concerts. Therefore, this paper focuses on model construction methods for different feature values.

For ticket purchase frequency, this paper employs the exponential smoothing method in time series analysis to detect frequency anomalies. Hence, the Holt-Winters seasonal decomposition model is selected. The specific process of establishing this model involves using the Holt-Winters seasonal decomposition method to decompose the time series into trend, seasonality, and residual components. Historical data is then used to train the Holt-Winters model and estimate parameters for trend, seasonality, and residual components. Subsequently, the ticket purchase frequency for each time window is calculated and compared with the model's predicted values. If the ticket purchase frequency for a specific time window significantly deviates from the model's predicted value, it is labeled as a frequency anomaly. The choice of the Holt-Winters seasonal decomposition model is based on its consideration of changes in trend, seasonality, and periodicity, allowing for more accurate capturing of data patterns. Additionally, by using exponential smoothing, it can better estimate changes in trends. Furthermore, it allows the selection of appropriate smoothing coefficients and cycle lengths based on actual circumstances, enhancing prediction accuracy. This model is widely applied in time series forecasting across various domains, demonstrating high predictive precision and accuracy.

For ticket purchase speed, this paper employs a regression model within supervised learning to predict the time taken for each ticket purchase stage and checks if it exceeds a predefined threshold. Regarding model selection, this study opts for a Support Vector Machine (SVM) regression model. The preparation of ticket purchase speed data includes the time taken by each user for each ticket purchase stage and labels indicating whether the ticket purchase is abnormal. The SVM regression model is trained using a training dataset, with the time taken for each ticket purchase stage serving as the target variable. The trained model is then used to predict the time taken for ticket purchase stages in a test dataset. Residuals, calculated as the difference between predicted and actual time, are compared to a predefined threshold. If the residual exceeds the threshold, the corresponding ticket purchase stage is labeled as abnormal. The rationale behind choosing this model lies in its strong generalization ability, capacity to handle high-dimensional data, determination of the decision function based on only a few support vectors, computational complexity dependent on the number of support vectors rather than the dimensionality of the sample space—thereby mitigating the “curse of dimensionality.” SVMs can handle nonlinear problems, identify crucial samples vital to the task, and exhibit robustness and interpretability.

For ticket purchase quantity, this paper utilizes an anomaly detection algorithm, setting an anomaly threshold to identify anomalies in ticket purchase quantity. The Isolation Forest model is chosen for this purpose. The model is trained using a training set, and then, the trained model is employed to detect anomalies in ticket purchase quantity in the test set. The Isolation Forest model returns anomaly scores for each ticket purchase quantity, and anomalies are flagged based on these scores and a predefined anomaly threshold. The reason for selecting this model lies in its effectiveness in identifying isolated anomalies without the need for distance or density metrics, leading to significant speed improvement and reduced system overhead.

Concerning payment methods, this paper employs a Random Forest classification model within supervised learning to predict whether the payment method is anomalous. Firstly, features for training the model are determined, which may involve payment time, payment amount, payment type, etc. These features are extracted and appropriately encoded. The Random Forest classification model is trained using the training set. The rationale for choosing this model is that it can handle high-dimensional data without the need for feature selection. It exhibits fast training speeds, is easily parallelizable, and can handle missing and outlier values. In situations of class imbalance, Random Forest provides an effective method for balancing dataset errors. Moreover, it outputs the importance of each feature, which can be utilized for feature engineering.

For IP address changes, this paper employs the K-Means clustering model from cluster analysis to detect abnormal patterns in IP address changes. The patterns of IP address changes are clustered, and anomalous clusters are labeled. The frequency of IP address changes is standardized to ensure uniform scaling. The standardized data are then trained using the K-Means clustering algorithm. K-Means clustering divides the data into K clusters. After training, each data point is assigned a cluster label. Data points that are distant from the cluster center are considered anomalies. The choice of this model is due to its simplicity, ease of implementation, applicability to large-scale datasets, fast computation, and suitability for various types of datasets.

For historical ticket purchase records, this paper employs a Long Short-Term Memory (LSTM) model from supervised learning to detect inconsistencies between historical behavior and new ticket purchase behavior. The ticket purchase history data are transformed into a time series data format suitable for LSTM. The data are then split into input sequences and target sequences, creating sample sequences for training. The input data are standardized or normalized to ensure consistency in scale. Finally, an LSTM model is constructed using the TensorFlow deep learning library. The reason for selecting this model is its ability to address issues such as gradient vanishing and exploding. It can handle long sequence data, introduces gate mechanisms to autonomously select which information to retain or forget, and effectively enhances the model’s expressive power.

Finally, concerning model integration, by progressively implementing the proposed models, they can be combined using a Voting method to merge models constructed for different feature values.

5. Conclusion and Outlook

This paper emphasizes the importance of anomaly detection in concert ticket purchase behavior, particularly as technological advancements lead to a greater diversity of anomalous behaviors. We have provided corresponding machine learning models for different feature values, laying the foundation for the construction of a comprehensive anomaly detection system. Of course, these models will require further fine-tuning of reasonable thresholds and parameters to ensure accurate detection of anomalous ticket purchasing behavior.

This study provides a crucial reference for the security and user experience of ticketing systems. Future research can expand in several directions. Firstly, we can further implement and optimize the models, which may involve additional feature engineering, fine-tuning of models, and the utilization of large-scale datasets.

Secondly, with the continuous development of deep learning and neural networks, consideration can be given to applying the latest technologies in this field to anomaly detection. Additionally, exploring the integration of real-time data streams into the anomaly detection system can enable faster responses to potential anomalous behavior.

Finally, attention should be consistently directed toward the evolution of ticket purchasing behavior and the emergence of new anomaly methods. Continuous improvement and upgrades to the anomaly detection system will ensure its adaptability to evolving challenges. This will guarantee that ticketing systems maintain a high level of security and user satisfaction while providing an enhanced ticket-buying experience for future concerts and large-scale events.

Author Contributions

The authors contributed to this work in the following ways:

- Zeng Lingyang: Conceptualization, Methodology, Writing – Half of the Original Draft, Review & Editing.
- Peng Lilin: Investigation, Data Curation, Writing - Half of the Original Draft, Review & Editing.

Both authors contributed equally to this work.

References

[1] Zhu, W., Li, W., Dorsey, E. R., et al. (2023). Unsupervised anomaly detection by densely contrastive learning for time series data. Neural Networks, 168, 450-458.

[2] Fan, J., Liu, Z., Wu, H., et al. (2023). Luad: A lightweight unsupervised anomaly detection scheme for multivariate time series data. Neurocomputing, 557, 126644.

[3] Wang, K. K. (2022). Research on abnormal behavior detection method of process data based on supervised learning (Doctoral dissertation, Shenyang Ligong University). DOI:10.27323/d.cnki.gsgyc.2021.000221.

[4] Fosić, I., Žagar, D., Grgić, K., et al. (2023). Anomaly detection in NetFlow network traffic using supervised machine learning algorithms. Journal of Industrial Information Integration. DOI:10.1016/j.jii.2023.100466.

[5] Song, X. L., Zhang, Y. B., Zhang, P. Y. (2022). Student consumption data anomaly detection based on semi-supervised learning. Computer and Modernization, 2022(12), 13-17.

[6] Zhou, L. J., Yan, Z. Y., Dai, L. L. (2019). Research and application of abnormal behavior classification technology in railway Internet ticketing. China Railway Science, 40(06), 133-139.

[7] Li, W., Zhu, J. S., Shan, X. H. (2018). Research and implementation of intelligent identification of abnormal users in railway Internet ticketing based on exponential weight algorithm. Railway Computer Application, 27(10), 7-10.

[8] Yang, L., Wang, F., Wang, T. (2017). Analysis of dishonorable behavior on railway online ticketing system based on k-means and FP-growth. 2017 IEEE International Conference on Information and Automation (ICIA), 1173-1177. DOI:10.1109/ICInfA.2017.8079079.

[9] Peng, Y., Yang, P. T., Song, J., et al. (2023). Research on user anomaly detection based on SVM under time sequence. Information Technology and Informatization, 2023(01), 130-133+138.