Machine Learning-Based Customer Churn Risk Prediction for Live Streaming e-Commerce

Xiangchen Meng, Runyu Li and Qianqian Song
Weifang Engineering Vocational College, Shandong Province, 262500, China

Keywords: Machine Learning, Live E-Commerce, Customer Churn

Abstract: As an emerging sales model, live streaming e-commerce has developed rapidly in recent years and has rapidly attracted many consumers. According to the data, the scale of the live broadcast e-commerce market in 2020 has reached more than 900 billion yuan. This impressive number has attracted many e-commerce companies, who have begun to explore and integrate into this live streaming sales field. For example, at 8 p.m. on March 20, 2020, Taobao Live's top streamer Wei Ya sold 560 million items during the live broadcast, attracting more than 700 million viewers, which significantly demonstrates the potential of live streaming to bring goods. In order to further promote the development of live e-commerce, Taobao Live announced on June 15, 2020 that it would invest a large amount of traffic resources to support live streamers on Taobao within a year, and launched a "new infrastructure" plan to upgrade Taobao's content ecosystem. Although live-streaming e-commerce has brought huge business opportunities, its rapid development has also raised some questions. First of all, the promotional strategies in the live broadcast room, such as "routines", "low prices" and "coupons", make it difficult for consumers to recognize. Secondly, the products promoted by the anchor may not match the actual needs of consumers. What's more, livestreaming e-commerce may be at risk of losing customers. To identify and predict this potential risk of churn, big data and machine learning techniques can be leveraged for analysis. Through these technologies, it is possible to gain insight into consumer behavior patterns and preferences, so that we can better anticipate and respond to possible market changes.

1 INTRODUCTION

In recent years, the mobile Internet has developed rapidly, which has also led to the rapid growth of the number of online shopping users in China. According to statistics, the number of online shopping users in China has reached 710 million in 2019, an increase of 12.2% compared with the previous year. In addition, with the widespread adoption of smartphones and the advent of the 5G era, there is still a lot of room for growth in the number of online shopping users. From 8.7 trillion yuan in 2013 to 20.4 trillion yuan in 2019, the scale of the online shopping market has expanded at an astonishing rate, with a compound annual growth rate of 18.6%. In addition, data shows that in 2019, the scale of China's live broadcast e-commerce market has exceeded 900 billion yuan, an increase of 205% over 2018. In 2020, affected by the epidemic, the live broadcast e-commerce industry entered a period of rapid development, and its market size is expected to reach about 11 trillion yuan by 2021. Consumers are paying more and more attention to the protection of their rights and interests and shopping experience (Chen, Hu, et al. 2023). For e-commerce companies, attracting users to spend and retain users has become a key issue. Especially for some traditional e-commerce, it is even more difficult to improve customer retention (Joardar Bletsch, et al. 2023). Because most traditional e-commerce uses products as the core selling point to attract users, while online shopping is user-centric to attract users (Jose, and Hampp, 2024). Therefore, in order to get more users to buy goods, it is necessary to improve the retention rate to retain users. Customer retention (LTV) is the ratio of revenue streams to total sales revenue over the lifetime of a customer, from the time they enter the website to the time they leave the website, (often used to measure customer lifetime value) (Kamoona, Song, et al. 2023). It can not only reflect the attractiveness of the company's products and services, but also measure the degree of attention the business has to customers. In the e-commerce industry, customer retention will directly affect the development of e-commerce companies in the future (Pontillo, Aragona, et al. 2024). Therefore, this paper chooses to analyze the problems of e-commerce

Meng, X., Li, R. and Song, Q.
Machine Learning-Based Customer Churn Risk Prediction for Live Streaming e-Commerce.
DOI: 10.5220/0013540000004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 1*, pages 283-288
ISBN: 978-989-758-763-4

enterprise customer retention rate and proposes solutions to improve customer retention rate (Schmidt, Kabir, et al. 2022). With the development of Internet technology and online shopping models, the traditional e-commerce industry has gradually begun to transform and upgrade to digitalization, intelligence and automation. In this process, big data and machine learning technologies have gradually been integrated into the e-commerce industry and play an important role. Both big data and machine learning can help companies identify and predict potential customer churn risks. That's why big data and machine learning techniques are used in this article (Swamy, 2022). This paper takes a live streaming e-commerce platform as the research object, firstly collects the user data, related product data, marketing campaign data and user purchase behavior data of the live streaming e-commerce platform as the data source, then uses the Spark framework to preprocess the relevant data, and finally uses random forest algorithm and XGBoost algorithm to construct a customer churn risk prediction model (Swamy, 2022).

2 RESEARCH METHODS

In this paper, three classification methods in the field of machine learning, including decision tree, support vector machine and random forest, are used to conduct predictive analysis based on live broadcast e-commerce data (Thipparaju, Sushmitha et al. 2022). The preprocessing step of the data includes dividing the dataset into a training set and a test set, and labeling the samples. The purpose of the normalization step is to optimize the model training performance. When building a model, you need to adjust the parameters to balance the accuracy of the model with the computational complexity. In this study, in order to facilitate further research, a decision tree algorithm was selected to construct a customer churn risk prediction model (Yang, Lee, et al. 2022). Decision trees are favored because of their simple structure, wide applicability, and robustness, while they have relatively low requirements for data preprocessing. The performance of the model is measured by classification accuracy, including area under the ROC curve (AUC), AUC, and mean absolute error (MAE). In general, the closer the AUC value is to 1, the better the classification performance of the model, and the smaller the MAE value, the better the model performance.

3 RESEARCH PROCESS

When processing the huge dataset of live streaming e-commerce, a large amount of unstructured data needs to be preprocessed, including data cleaning, feature refinement, and data standardization. The detailed steps are as follows: First, data cleaning is carried out, and according to the characteristics of live broadcast e-commerce, the user's purchase records should be carefully cleaned, and abnormal data and duplicate records should be eliminated to improve the accuracy of data processing. After that, feature extraction is carried out to classify user attributes, such as gender, age, watch time, etc., and use these attributes as features for subsequent modeling. After feature extraction is complete, a predictive model needs to be built to predict the purchase behavior of new users and prevent customer churn. Customer churn risk prediction is divided into two categories: classification and regression: for classification problems, random forest algorithms are used to predict, such as XGBoost, XPadding, XGBoost+XGBoost and other models, to predict whether new users will buy products. For the regression problem, the logistic regression algorithm is used to make predictions. In view of the large number of attributes and large amount of data in live broadcast e-commerce, the logistic regression algorithm is prone to overfitting, so the K-means algorithm is used to classify the data to reduce the risk of overfitting. Finally, the performance of the model needs to be evaluated by cross-validation, including indicators such as accuracy, recall, and F1 value. Based on the results of the comprehensive evaluation, the optimal model is selected and new users are recommended.。

3.1 Data Collection and Pre-Processing

In the field of live streaming e-commerce, user data plays a crucial role. However, in actual operation, there are often defects in the collection and preprocessing of data, resulting in incomplete and non-standard data, which brings challenges to the establishment of models. In view of this, the beginning of the work should focus on the cleaning and sorting of data, and build a model on this basis. In this study, we used R language to analyze the user data of live streaming e-commerce, and selected user browsing history as the main data input. In view of the fact that users' browsing history on live streaming e-commerce platforms is recorded on an hourly rather than daily basis, the study adopted an innovative approach to data cleaning and processing. The specific steps include: first, classify the user's browsing history according to the time period, distinguish between the behavior within a fixed time period (such as "watching live streams" and "normal browsing") and the behavior outside the fixed time period (such as "purchase" and "favorite"), and then calculate the proportion of each behavior in a fixed time period by calculating its relative frequency in a fixed time period. The specific method is described as follows: 1. Calculate the total number of times the user's browsing history in the specified time range, and allocate these total times into five equal-length time periods, each of which contains 100 browsing records. 2. Normalize the number of times a user browses their history in each time period to eliminate the impact of differences in the number of views in different time periods. 3. Compare the proportion of each behavior in each time period with the proportion of the overall fixed time period to analyze the stability of user behavior patterns. Through these processing steps, we can observe that the relative weight of the user's browsing history is stable in each time period, which reveals that the user's behavior pattern has a high consistency over a fixed period of time. This finding can help to better understand user behavior and provide more effective user analysis and strategy support for live streaming e-commerce.

3.2 Feature Selection and Construction

In order to make live e-commerce operate better, it is necessary to predict the risk of customer churn, and in this process, it is necessary to extract the relevant factors that affect the risk of customer churn and take these factors as characteristics, so as to establish a prediction model. Features need to be selected and selected before a model can be built, and this is because data is the foundation of machine learning. If the data contains too many features, it can cause overfitting. So how to select features? This needs to be considered from three aspects: 1. Evaluate the importance of features. 2. The degree to which the calculation features influence the classification results. The importance can be evaluated based on different criteria, such as information gain-based and information-based entropy, or the distance between the calculated feature and the predicted result, or the distance between the calculated feature and the label. Secondly, for live broadcast e-commerce, live broadcast products are the medium of transaction between the anchor and the user, rather than the transaction object. Therefore, for live broadcast e-commerce, the selection of live broadcast products pays more attention to the trust between the anchor and the user, which requires the selection of live broadcast products with a high degree of trust. When selecting features, you can also consider selecting live broadcast products based on the customer's trust in the anchor, such as selecting live broadcast products based on the customer's trust in the anchor, and selecting live broadcast products according to the customer's trust in the anchor. When building a model, you need to determine the model type and parameters according to different situations. For supervised learning, linear models or support vector machines can be used as prediction models, and decision trees or naive Bayes can be used as prediction models for unsupervised learning.

3.3 Selection and Construction of Machine Learning Models

In this study, we use decision tree as the main machine learning method, combined with K-means clustering algorithm to reduce data dimensionality. As a supervised learning algorithm, decision trees are suitable for binary or multivariate classification problems. The workflow consists of separating all training samples into a training set and a test set, training the model on the training set, and using the results of the test set to evaluate whether the model performs as expected. In the process of building a decision tree, we will use concepts such as information entropy, information gain, etc. Information entropy is a metric used to assess the degree of disorder of a dataset, and it is calculated as follows:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \; \log_2 P(x_i) \tag{1}$$

where is the $P(x_i)$probability that the value of the random variable x is valued. x_i

The K-means clustering algorithm is an unsupervised learning method, which sorts the attribute values of the samples in the dataset and clusters them according to certain weights, and finally divides the data into multiple clusters. The core of the algorithm is to iteratively optimize the center point of each cluster until the convergence condition is reached. Let the dataset be and the number of clusters is$D = \{x_1, x_2, \ldots, x_n\}$ k, then the goal of K-means clustering is to minimize the sum of squares of the distance from the sample point in the cluster to the center point of the cluster, i.e.,

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \mid x_i - \mu_k \mid^2 \tag{2}$$

where the $C_k$first cluster is denoted and the center point of the first cluster is described. $k\mu_k k$

In the scenario of churn risk prediction of live broadcast e-commerce customers in this study, the K-means clustering method is first used to reduce the dimension of the dataset. The specific method is to sort each record in the dataset according to its attribute value, and then use the K-means algorithm to group the sorted data to form a high-dimensional feature set. Next, the training set and test set were used to build the prediction model, and the classification algorithm, regression algorithm and decision tree algorithm were used in the construction process. In particular, in the training stage of the decision tree model, the optimal segmentation attribute is selected by comparing the information gain, gain rate, or Gini index. The evaluation of the model predicts the test set data through the trained decision tree, and uses the accuracy, recall, F1 value and other indicators to evaluate the performance of the prediction results.

3.4 Model Training and Parameter Optimization

In this study, the decision tree model is our core technology. Segmentation of the data is the first step in order to distinguish between the training set used to build the model and the test set used to evaluate the model's performance. After the architecture and hyperparameters of the model are set, we further divide the training data into a training set and a test set according to certain rules. This process needs to be carried out under consistent conditions to guarantee reliable results. For the training data, we set preliminary hyperparameter values based on past experience. After the hyperparameters have been tuned and optimized, we calculate the corresponding error rate, which helps us to select the best combination of hyperparameters. The calculation of the error rate relies on a specific formula that involves factors such as prediction error, classification standard deviation, and confidence level.

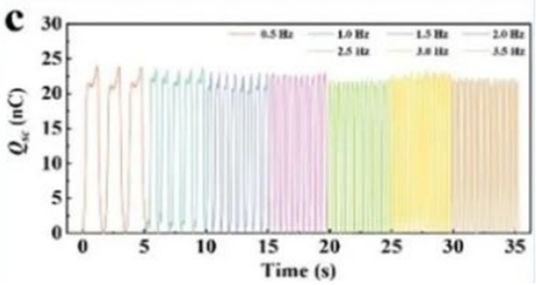

Figure 1: Machine learning-based customer churn risk

Once the optimal hyperparameters are determined, we use optimization algorithms to further improve the performance of the model. Under the supervised learning framework, we train the model by constructing a mapping between features and labels. By carefully selecting the hyperparameters, we are able to effectively map features to the label space, with the goal of improving the accuracy of predictions. In this study, we used the Least Squares Support Vector Machine (LSSVM) model and the Genetic Algorithm-based Optimization Strategy (GA-BP) to optimize this mapping process. To evaluate the performance of the model under different hyperparameter configurations, we used evaluation metrics such as accuracy (AUC), precision (MAPE), recall (RE), and F1 value to comprehensively evaluate the performance of the model.

3.5 Evaluation and Validation of the Model

Several effective methods are commonly used in the academic community to evaluate the performance of prediction models, including ROC curve analysis, AUC value evaluation, and Fisher's Information Coefficient (FICI). The ROC curve is plotted by connecting the points of the true positive rate (TPR) and false positive rate (FPR) at different thresholds to form a curve, which allows for a visual evaluation of the model's performance. However, when the amount of data is large, the effectiveness of this approach may be reduced.

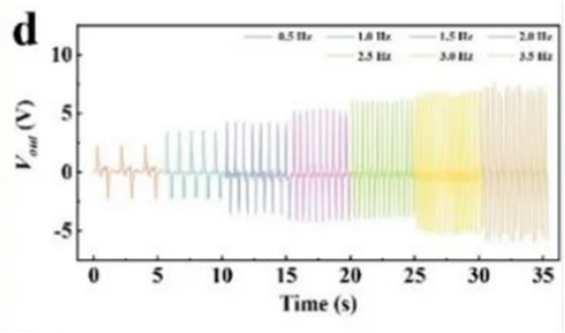

Figure 2: Prediction for live streaming e-commerce

The AUC value is judged by comparing the predicted results of the model with the actual results. Fisher's information coefficient rule measures the importance of features by calculating their FICI values, with higher FICI values indicating better predictions of the model and worse conversely. For specific datasets, the most appropriate evaluation method can be selected according to the characteristics of the data to optimize the prediction accuracy of the model. Typically, the evaluation of the model is also performed by calculating the AUC value under the ROC curve, i.e., by calculating the average of the AUC values of the training set and the test set to evaluate the overall performance of the model. The AUC value is used as the evaluation criterion for the performance of the model, which can effectively judge the prediction accuracy of the model.

4 FINDINGS:

In the data preprocessing stage, the data were normalized and normalized, and the features were divided into two categories, 0 and 1, according to the eigenvalues, and the results were stored in the variable labels. In order to construct the model, the original dataset was first normalized, and then the random forest algorithm was used for feature selection, and the model was trained based on the training set. In addition, the test dataset was divided into a training set, a test set, and a cross-validation set for model accuracy validation and generalization ability evaluation, respectively. During the model training process, 22 features were selected from the training set as training data and these data were standardized. The operation mechanism of the random forest algorithm includes: randomly selecting the input and output features, generating the output based on a series of decision rules, calculating the weight of each rule, and finally selecting the output result of the rule with the largest weight, and comparing it with the original data to judge the correctness. In the model building stage, five algorithms, including logistic regression, random forest, support vector machine, linear regression, and naive Bayes, are used to train the model. The comparison results show that the random forest algorithm has the highest prediction accuracy of 94.12%, followed by logistic regression and support vector machine, while the linear regression has a lower accuracy of 83.10%, and the accuracy of random forest is 2.0% higher than that of the original data. Further analysis of the test set shows that the highest classification accuracy is 98.23%, which is achieved by the linear regression model, and the support vector machine and random forest have the best accuracy, which are 94.10% and 83.10%, respectively. This result shows that random forest has significant advantages in classification tasks, which is suitable for predicting the churn risk of live streaming e-commerce. Through the comparison and analysis with other algorithms, the relationship between prediction accuracy and effect is found, and the algorithm performance is discussed in depth.

5 CONCLUSIONS

In this study, the K-Means method is used to segment the live streaming e-commerce user group, and then the three machine learning techniques of decision tree, random forest and support vector machine are combined to estimate the churn probability of users. It is found that in terms of accuracy, random forest performs the most outstandingly, with an accuracy rate of 86.34%, while the support vector machine has a relatively low accuracy of 64.75%. In order to make a more effective prediction of the churn of live streaming e-commerce users, the ROC curve and AUC value are introduced as evaluation tools. The ROC curve can intuitively reflect the predictive performance of the model, and the AUC value provides a concise evaluation standard. Based on these analyses, this paper uses the random forest algorithm to classify users, and uses the ROC curve to evaluate the classification effect to predict whether users will leave the live streaming e-commerce platform. Based on the experimental data, the following insights are obtained: First, the random forest algorithm provides the highest prediction accuracy in the churn prediction task of live broadcast e-commerce. Secondly, the churn of users is closely related to purchase behavior, which is the basis for our selection of these algorithms for research.

Thirdly, the ROC curve shows that the random forest algorithm can not only accurately predict user churn, but also effectively evaluate the performance of the model. Finally, we suggest some potential areas for improvement: introduce the "churn level" metric and use it to determine if further service actions are needed for users.

REFERENCES

Chen, G., Hu, Q. C., Wang, J., Wang, X., & Zhu, Y. Y. (2023). Machine-learning-based electric power forecasting. Sustainability, 15(14)

Joardar, B. K., Bletsch, T. K., & Chakrabarty, K. (2023). Machine learning-based rowhammer mitigation. Ieee Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42(5), 1393-1405.

Jose, B., & Hampp, F. (2024). Machine learning based spray process quantification. International Journal of Multiphase Flow, 172

Kamoona, A., Song, H., Keshavarzian, K., Levy, K., Jalili, M., Wilkinson, R., . . . Meegahapola, L. (2023). Machine learning based energy demand prediction. Energy Reports, 9, 171-176.

Pontillo, V., d'Aragona, D. A., Pecorelli, F., Di Nucci, D., Ferrucci, F., & Palomba, F. (2024). Machine learning-based test smell detection. Empirical Software Engineering, 29(2)

Schmidt, A., Ul Kabir, M. W., & Hoque, M. T. (2022). Machine learning based restaurant sales forecasting. Machine Learning and Knowledge Extraction, 4(1), 105-130.

Swamy, G. (2022a). Machine learning based face recognition system. International Journal of Early Childhood Special Education, 14(3), 4999-5006.

Swamy, G. (2022b). Machine learning based face recognition system. International Journal of Early Childhood Special Education, 14(3), 4775-4782.

Thipparaju, R., Sushmitha, R., & Jagadeesh, B. N. (2022). Machine learning based renewable energy systems. International Journal of Early Childhood Special Education, 14(3), 2197-2203.

Yang, Q. D., Lee, C. Y., Tippett, M. K., Chavas, D. R., & Knutson, T. R. (2022). Machine learning-based hurricane wind reconstruction. Weather and Forecasting, 37(4), 477-493.