Customer Churn Prediction based on E-commerce Live Streaming Data

Mukun Chang

Student ID: 557607

Thesis supervisor: Dr. Kathrin Gruber

Second assessor: Prof. Dr. Bas Donkers

Master thesis

Data Science and Marketing Analytics

Erasmus School of Economics

Erasmus University Rotterdam

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Date: August 2, 2023

Abstract

In 2016, the new marketing form of "online live streaming + e-commerce" began to enter the mainstream life. With the promotion of the "stay-at-home economy" during the epidemic, e-commerce live streaming has entered a rapid development stage worldwide. Different from traditional online shopping, e-commerce live streaming possesses the characteristics of real-time interaction, providing consumers with efficient and timely shopping experiences. For platforms and live streaming operators, e-commerce live streaming can reduce marketing costs, improve sales efficiency, and enhance trust between them and consumers. However, as a new form of e-commerce, customer churn must be taken into consideration. Identifying the key factors influencing customer retention and predicting potential customer churn through statistical models can have a significant impact on live streaming platforms. Data mining models are well-suited to meet this requirement. This paper utilizes customer data collected from e-commerce live streaming platforms and applies unsupervised machine learning models to classify customers, understand their current status and characteristics, and perform churn prediction analysis. By preprocessing behavioral features and examining the coefficients of a logistic regression model, the core traits of different customer segments are identified, thereby enhancing the practical value of the model's results.

Contents

1 Introduction

1.1 Research Background

E-commerce live streaming is a shopping method that falls under the category of commercial advertising activities in the legal domain. Hosts are responsible for specific behaviors as "advertising endorsers," "advertising publishers," or "advertisers." Viewers purchase goods by placing orders within live streaming rooms. As a form of socialized e-commerce, live-streaming e-commerce possesses a stronger sense of consumer trust and engagement (Wongkitrungrueng and Assarut, 2020).

When it comes to live-streaming e-commerce, it is impossible not to mention China, as it first emerged in China in 2016. With the widespread adoption and development of internet technology, China has amassed a large number of internet users, known as netizens. Each netizen has the potential to become a prospective user of live-streaming e-commerce. According to data released by the China Internet Network Information Center (2023), as of December 2022, the number of internet users in China reached 1.067 billion. The scale of online live-streaming users in China reached 751 million, among which the scale of e-commerce live-streaming users was 515 million, showing an increase of 51.05 million compared to December 2021. With such a massive user base, it is evident that live-streaming e-commerce is immensely popular in China. However, live-streaming e-commerce is not limited to China alone. To adapt to the changing landscape of the retail industry, many retailers and brands have started exploring e-commerce live-streaming on platforms such as Twitch, Instagram, and TikTok, including companies like Walmart, Nordstorm, and Burberry (Kumarab & Venkatesanc, 2021). Liu (2020) pointed out that due to the impact of the COVID-19 pandemic, the influencer economy and live-streaming e-commerce industry came into the public eye after December 2019. Scholars worldwide have gradually started paying attention to live-streaming e-commerce, which is exceptionally popular in China but relatively less prominent in other parts of the world. They have begun conducting in-depth research on live-streaming e-commerce, providing diverse insights and references for global enterprises and media. Cunningham et al.'s (2019) study states, "Compared to the West, live-streaming industry practitioners in China have more opportunities and platforms. E-commerce live-streaming has led e-commerce platforms towards multi-channel integrated marketing, driving the development of emerging consumer culture in China. Their success also indicates that in the future, the West may achieve similar success as we also possess the technologies they employ."

1.2 Research Objectives and Significance

Considering that live-streaming e-commerce started rencently, research in related fields is also lagging behind. Most studies are trend-oriented and lack in-depth analysis and exploration. Live-streaming e-commerce plays a positive role in economic development and business channel expansion, thus necessitating a better understanding of this e-commerce form. With the continuous improvement of network infrastructure, the iterative advancements in internet technology, and the widespread use of various mobile smart devices, live streaming e-commerce has been steadily growing. As the number of users engaging in shopping through live streaming continues to increase, the abundance of product and user information has led to the issue of information overload. For users, the extensive variety of products makes it difficult to make choices. For live streaming platforms, analyzing users within vast amounts of data and achieving predictions of users' live shopping behavior has become a challenging problem. Live streaming operators and broadcasters consistently face the challenge of customer churn. It becomes crucial for them to summarize the current status and behavioral characteristics of customers based on backend data, as this can assist in implementing diverse marketing strategies for potentially churned customers.

Against this backdrop, this study will focus on researching the prediction of live streaming users' shopping behavior based on order data from a specific electronic product live streaming room in China. By utilizing machine learning to predict customer retention based on orders placed within the live streaming room and studying the trained model, we can analyze live streaming users' shopping behavior. This thesis employs an empirical analysis approach. The product under investigation in this paper is live streaming e-commerce. Compared to other forms of sales, the user structure in this context is more complex. To effectively differentiate users, it is necessary to incorporate clustering models. Additionally, since the cost for users in live streaming e-commerce is almost negligible and user loyalty is low, the window of opportunity for platform retention of potential churners is fleeting. Therefore, apart from accurately predicting user churn, it is crucial to study users' core behaviors and understand their retention motivations in order to optimize product and algorithm recommendations. The specific steps designed for the empirical analysis of this study are as follows:

(1) Feature selection and data preprocessing;

(2) Application of clustering models for user clustering and interpretation of different clusters;

(3) Utilization of logistic regression models for feature importance analysis, identifying core factors and providing interpretations;

(4) Application of logistic regression, random forest, and gradient boosting decision tree (GBDT) models for churn prediction analysis;

(5) Combining insights gained, providing explanations and recommendations for product and algorithm enhancements.

Firstly, this study holds strong academic relevance. As mentioned earlier, research on e-commerce live streaming is still in its nascent stages, with no prior focus on customer churn in e-commerce live streaming and the application of machine learning in studying customer churn in this context. In the subsequent literature review section, I will provide a more detailed discussion on existing research and its limitations. Secondly, in terms of managerial relevance, both live streaming operators and individual broadcasters can benefit significantly from this study. It can help them identify potential churned users in a timely manner and develop targeted retention strategies that align with the core needs of users, thereby enabling precision marketing. Furthermore, based on user behavior, classifying users can help operators understand the current status and behavioral characteristics of different user segments, facilitating subsequent fine-tuning of strategies and operations for each user segment. Lastly, live-streaming e-commerce is a form of e-commerce, which lends societal relevance to this study. Undoubtedly, in the post-pandemic era, governments around the world are making efforts to stimulate consumption. Effective precision marketing by live streaming operators can contribute to expanding domestic demand in various countries.

1.3 Research Framework

This thesis is divided into six sections, organized as follows. Section 1 provides an introduction to the research background, presents a summary of the research's importance, outlines the research content and methods, proposes potential innovative points, and clarifies the overall direction for further in-depth investigation. Section 2 provides a summary and analysis of the current research in the field of live-streaming e-commerce, as well as studies on machine learning-based customer churn and precision marketing. Section 3 will provide an introduction to the dataset used for the analysis and the data preprocessing conducted for subsequent analysis. Section 4 will present the relevant theoretical foundations that will be utilized in this thesis. Section 5 will present the empirical analysis results and their interpretations. Section 6 will propose the conclusion, recommendations, and future prospects.

2 Literature

This section aims to summarize, organize, and analyze existing research achievements in two aspects: e-commerce live streaming and customer churn and precision marketing based on machine learning.

2.1 E-commerce Live Streaming

From an IT perspective, Sun and Li (2019) defined e-commerce live streaming as an information dissemination medium that involves extensive human-computer interaction processes. In simple terms, e-commerce live streaming refers to sellers communicating directly with consumers through live broadcasts, enabling consumers to place orders within the same system (Wang et al., 2022). Compared to traditional e-commerce, live streaming e-commerce possesses the following advantages.

(1) *Authenticity*

Cui et al. (2023) argue that the advantages of e-commerce live streaming lie in the ability of hosts to interact in real-time with online viewers while introducing, showcasing, and selling products through digital devices. Chang et al. (2015) also point out that consumers in live streaming rooms can gain a more authentic, prompt, and comprehensive understanding of product information. According to Chen et al. (2017), their study revealed that hosts in the context of e-commerce live streaming can offer effective product demonstrations and immersive consumer experiences, thereby positively influencing consumers' purchase intentions;

(2) *Celebrity effect*

Geng et al. (2020) put forward that with the widespread popularity of social media and live streaming, influencer endorsements have become a popular content marketing approach for e-commerce sellers. Meng et al. (2021) indicate that the essence of e-commerce live streaming is social commerce, and the participation of celebrities and internet influencers significantly motivates viewers to make purchases. Zhu and Li (2021) conducted a study using an adjusted chain-mediation model and found mutual influences between consumers' heightened interactive enthusiasm and the attractiveness of hosts during live broadcasts;

(3) *Entertainment value*

Due to the multi-entity interactive features within live streaming rooms, viewers easily enter a state of flow, where they ignore the presence of other stimuli, immerse themselves in the current activity, and are willing to invest a certain cost for the sake of happiness (Ha et al., 2007). Research has shown that flow experience has a positive impact on consumer loyalty and purchasing behavior (Ling et al., 2011). Based on the information foraging theory and multimedia learning theory, Wang and Wu (2019) conducted an analysis of the impact of consumer engagement mechanisms in e-commerce live streaming. The study revealed the significance of interactive entertainment in e-commerce live streaming for businesses to increase potential customers. Lu (2019) emphasizes that the evolution of media facilitates the innovation and development of marketing models. As a novel form of communication media, online live streaming attracts a large customer base through real-time interaction and comprehensive audiovisual experiences.

Despite the aforementioned advantages, e-commerce live streaming also faces numerous challenges. In comparison to platform and host evaluations of products, customer evaluations hold greater influence (Bickart and Schindler, 2001). Singh et al.'s (2017) research indirectly indicates that the factors influencing viewers' purchasing intentions and behaviors in live streaming rooms are diverse and complex. Currently, research on e-commerce live streaming primarily focuses on three aspects: the willingness of hosts to use live streaming platforms, the willingness of viewers to watch live streams, and the willingness of viewers to make purchases within live streaming rooms (Hu et al., 2017). There is a clear lack of research specifically focused on customer churn prediction in e-commerce live streaming.

2.2 Customer Churn and Precision Marketing Based on Machine Learning

With the development of the internet, research on user behavior prediction and customer churn has become a popular direction among scholars. As early as 1996, Steward conducted a study on churn among telecom users (Steward, 1996). Sladojevie et al. (2011) used data mining techniques to predict customer churn in telecommunications and gained insights from it. Gursoy (2010) applied logistic regression and decision trees, two classification techniques, to identify the reasons for customer churn in telecom companies. Kim et al. (2015) proposed a feature-based multiple regression model to predict telecom customer churn and provided recommendations for churn prevention. Bach et al. (2021) designed a three-stage approach to analyze customer churn in telecommunications, incorporating cluster analysis and decision trees. Bugajev et al. (2022) employed six different classification methods to obtain the highest accuracy in predicting customer churn in the telecom industry. Kim et al. (2015) used convolutional neural networks to predict customer churn after telephone marketing. Long et al. (2012) developed a predictive model based on cluster schemes to analyze potential churn among users in social networks and provided suggestions for retaining these users. Addressing customer churn in the manufacturing industry, Liu and Ju (2009) designed a support vector machine-based prediction model and employed component analysis for dimensionality reduction to gain insights. Summarizing research worldwide, it is evident that significant findings and achievements have been made in the application and analysis of customer churn prediction, particularly in the telecommunications industry. However, regarding live streaming platforms, especially in the context of e-commerce live streaming, there remains ample space for research and analysis in predicting customer churn.

Patrick (2001) explored the idea of providing enterprise products to the right customers to achieve low-cost marketing objectives. Research has indicated that big data analysis and machine learning play a crucial role in personalized sales, specifically in the areas of customer information acquisition and management, more precise market positioning, assisting in the development of marketing and product promotion strategies, and analyzing market environment and user predictive data (Chaudry et al., 2018). Alkhayrat et al. (2020) pointed out that utilizing machine learning in consumer profiling, strategy support, and enhancing customer experience can help companies excel in customer-centric sales and customer service. Companies can effectively employ machine learning for marketing purposes, such as analyzing commodities, users, channels, costs, and gaining insights into user needs to explore marketing methods for products (Yang and Zhang, 2018). Claudio et al. (2011) constructed two types of real estate price prediction models using model trees and multivariate adaptive regression splines techniques, aiming to accurately define segmented markets. Wang et al. (2018) designed a Spark-based recommendation system and proposed relevant algorithms and strategies to achieve a machine learning-based precision marketing system. Overall, the significance of precision marketing and its widespread application are evident. However, precision marketing in the field of live streaming e-commerce has not received much attention. This study can integrate churn analysis with precision marketing to provide a unique analytical approach for the live streaming e-commerce industry.

3 Data

3.1 Data Description

Compared to other regions around the world, live streaming e-commerce has experienced the most rapid growth in China. Taobao, founded by Alibaba Group in May 2003, is one of the largest online retail platforms in the Asia-Pacific region. In China, Taobao dominates nearly 80% of the live shopping market, while Douyin and Kuaishou (both are short video apps in China like Tiktok) share the remaining portion. Taobao's platform is the most comprehensive and holds the largest market share in China's e-commerce live streaming sector. Therefore, Taobao's live streaming data has strong practicality and representativeness for customer churn and precision marketing research.

The dataset used in this study is the backend data of a certain electronics brand's live streaming rooms on Taobao, which was publicly released on Alibaba's Tianchi Forum. The brand's main products are electronic devices, but it also sells other products including smart home appliances and fashion items. The backend data of customers who placed orders in this live streaming room in February 2022 was collected for the month of March (these customers have been identified as having watched the live broadcasts of the livestream room in March), and these data have been processed to ensure the protection of customer privacy and prevent any potential data breaches. All variables were collected based on the status of these customers as of the end of March 2022. The dataset is divided into the following dimensions:

(1) Demographic characteristics of customers (e.g., marital status, age);

(2) Usage patterns of the live streaming software (e.g., duration of app usage, preferred login devices);

(3) Customer purchasing behavior data (e.g., number of orders in the previous month, usage of discount coupons in the previous month);

(4) Customer account information (e.g., number of followers, account balance);

(5) Churn flag (If a customer made a purchase in this live streaming room in March, the value is 1; otherwise, it is 0).

3.2 Data Preprocessing

In order to improve the training and prediction effectiveness of the machine learning models, it is necessary to perform data preprocessing on the aforementioned dataset.

3.2.1 Missing Data Analysis

First, to understand the missingness of features in the dataset, the missing values in the dataset were visualized, as shown in Figure 1.

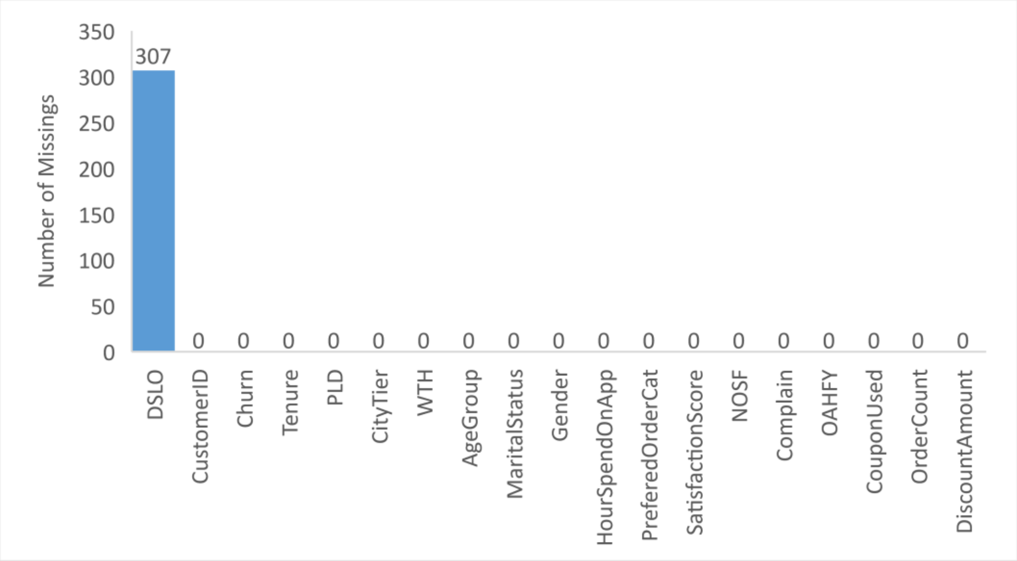

Figure 1. Distribution of missing values

As shown in the figure, it can be observed that all the missing values are concentrated in the variable “DSLO”, which represents the number of days since the customer's last entry into the live stream. We know that all customers watched the live stream in March, indicating that they accessed the live stream during that month. The data collection period ended on the last day of March, so this variable represents the number of days between the customers' last login in March and the last day of March. There is no clear correlation between customers' purchase intention and this variable, thus “DSLO” is removed from the dataset. “OAHFY” represents the growth rate of order quantity compared to the previous year. Since we aim to predict customer behavior in March based on their account status in February, “OAHFY” is not relevant to our analysis. Moreover, the customer ID numbers are not meaningful for our subsequent analysis; therefore, “CustomerID” is also removed.

3.2.2 Normalization Process

As each variable exhibits a distinct range of values, during the modeling process, factors with larger value ranges exert a more significant influence on the model, while factors with smaller value ranges may be disregarded. Furthermore, the varying scales among different factors require the elimination of the impact of these scale differences. Consequently, data standardization is necessary. Subsequently, by comparing the magnitudes of variable coefficients, we can interpret the importance of features.

3.2.3 Encoding of Non-numeric Features

However, before normalization, we observed that not all variables in the data were numeric; there were also some factors represented as categorical data. First, select the four non-numeric attributes from the dataset: “PreferredLoginDevice”(“PLD”) “MaritalStatus”, “Gender”, and “PreferedOrderCat” as shown in Table 1.

Variable name	Description	Type
PLD	Preferred login device of customer	Character; “Mobile phone” & “Phone”
MaritalStatus	Marital status of customer	Character; “Divorced”, “Married” & “Single”
Gender	Gender of customer	Character; “Male” & “Female”
PreferedOrderCat	Preferred order category of customer in last month	Character; “Fashion”, “Grocery”, “Household”, “Laptop & Accessory”, “Mobile Phone” & “Others”

Table 1. Non-numeric attributes

These four variables are all categorical variables. Therefore, the characters in these four variables will be converted into numbers based on their categories, and the data type will be converted to the factor type. The variable "CityTier" represents the level of the city, with levels ranging from one to three. In China, cities can be categorized into different tiers or lines, with first-tier cities being the most advanced, followed by second-tier cities, and lastly, third-tier cities. In this classification system, a lower numerical value indicates a higher level of urban development. I grouped the divorce and single categories together as one group, and the married category as another group, thereby converting this variable into two categories: unmarried and married. Regarding the variable "PreferedOrderCat," my approach was to categorize the products into two classes: 0 representing non-primary products and 1 representing primary products. As the data originated from a live streaming platform of an electronic product brand, computers and smartphones were considered primary products, while the remaining products were classified as non-primary products. The reasons for the above operations are twofold. Firstly, certain categorical variables are transformed into trend-like variables that can be used for regression analysis. For instance, the level of a city, where smaller numbers indicate relatively better city development. Secondly, while ensuring insightful analysis, some multi-categorical variables are controlled and converted into two-category variables. For example, product types are constrained to two categories: “main products” and “non-main products”. This simplifies the model and facilitates the interpretation of results.

3.3 Variables

After the aforementioned data preprocessing, the dataset consists of 56300 observations. Table 2 displays all the variables along with their meanings. All variables have been normalized. In the table, certain variables are only presented in their pre-normalized state, with the purpose of providing readers with a more intuitive view of the data preprocessing performed.

Variable name	Description	Type
Churn	Churn Flag	If a customer made a purchase in this live streaming room in March, the value is 1; otherwise, it is 0; Factor
Tenure	Tenure for using the platform	Numeric
PreferredLoginDevice (PLD)	Preferred login device of customer	Mobile Phone or Phone=0, Pad= 1 (Before normalization)
CityTier	City Tier	1 – 3 (Before normalization)
WarehouseToHome(WTH)	Distance in between warehouse to home of customer	Numeric
AgeGroup	The age of customer	Numeric
MaritalStatus	Marital status of customer	Unmarried=0, Married=1 (Before normalization)
Gender	Gender of customer	Male=1, Female=0 (Before normalization)
HourSpendOnApp	Number of hours spend on App last month	Numeric
PreferedOrderCat	Preferred order category of customer in last month	Non-primary=0, Primary=1 (Before normalization)
SatisfactionScore	Satisfactory score of customer on service	Numeric
NumberOfStreamerFollowed (NOSF)	Total number of streamers followed by particular customer	Numeric
Complain	Any complaint has been raised in l ast month	If a complaint has been raised, the value is 1; otherwise, it is 0 (Before normalization)
CouponUsed	Total number of coupon has been used in last month	Numeric
OrderCount	Total number of orders has been places in last month	Numeric
DiscountAmount	Average cashback in last month	Numeric

Table 2. Variables in the data set

4 Methodology

4.1 Cluster Analysis

Cluster analysis is a method that categorizes objects based on the similarity of their feature indicators. Throughout the process of clustering, it automatically groups targets by considering the overall differences among various features, rather than predefining an accurate classification criterion. Currently, there are numerous research achievements in the field of clustering methods, with the most commonly used approaches including partitioning methods, hierarchical methods, and density-based methods. In this thesis, we will focus on introducing the most classic density based algorithm and methods for determining the appropriate value of k.

4.1.1 K-means Clustering Algorithm

The k-means algorithm is a partitioning clustering algorithm where the Euclidean distance between points is used as a measure of similarity between different observations. In other words, by considering k as a parameter, the algorithm divides n data points into k clusters, aiming to achieve high similarity within each cluster and low similarity between clusters. In the case of the k-means clustering algorithm, if any point among the various points is relatively close to a specified point, then that point and the point with the shortest distance to the specified point are considered to belong to the same cluster.

The specific implementation process of the k-means clustering algorithm is as follows:

(1) Randomly select k initial centroids as desired clusters from the given dataset. The centroid represents the central point of all the points within a cluster, typically computed as the mean of these points.

(2) For the remaining data, calculate the distance between each data point and the centroids, and assign each data point to the centroid with the smallest distance.

(3) Recalculate the centroids of each cluster based on the data points assigned to it.

(4) Repeat steps 2 and 3 until each cluster no longer changes.

4.1.2 Determining the Value of k in k-means Clustering Algorithm

Elbow Method: The core idea of this method is the application of the sum of squared errors (SSE). SSE represents the clustering error of all samples and serves as an effective evaluation of the clustering performance. By plotting the relationship between k and SSE, the graph resembles an elbow, where the k value corresponding to the elbow indicates a significant convergence of SSE, which represents the optimal number of clusters for the data. This metaphorical term "elbow" accurately describes the rationale behind the name of this method.

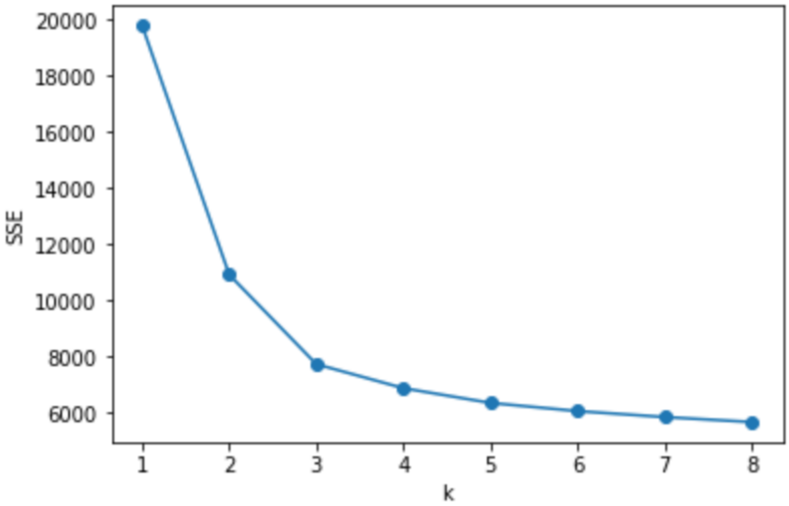

Figure 2. An Example: Graph of the Relationship between k and SSE

4.2 Logistic Regression

Logistic regression is a type of generalized linear regression model, which is essentially a linear classifier known as "regression." Despite its name, logistic regression is widely used in classification problems and is commonly applied in fields such as data mining and economic forecasting. To understand the origins of logistic regression, it is necessary to first comprehend linear regression. In machine learning, linear regression is the simplest regression algorithm, and its formulation is straightforward and easily understandable.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

β is collectively referred to as the parameters of the model, with β_0 being referred to as the intercept and β_1 to β_n as the coefficients.

This formula can be simplified as :

$$z = \sum_{i=0}^{n} \beta_i x_i = \beta^T x \tag{2}$$

The task of linear regression is to map the linear relationship between the input feature *x* and the label values *y* using a prediction function. Our aim is to find a parameter β such that the z function fits the data best. Least squares is often used in mathematical studies to solve for parameters in linear regression.

For discrete variables following a 0-1 distribution, we introduce a bridge to transform the linear regression equation z into g(z). Simultaneously, we ensure that the values of g(z) are distributed between (0, 1), and when g(z) approaches 0, the sample label is assigned to class 0. Similarly, when g(z) approaches 1, the sample label is assigned to class 1. This yields a classification model. For logistic regression, the bridge is the Sigmoid function:

$$g(z) = \frac{1}{1+e^{-z}} \tag{3}$$

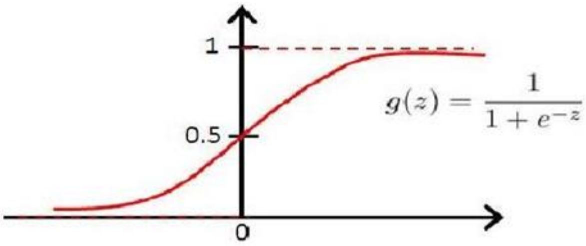

Figure 3. Sigmoid function graph

As shown in Figure 3, the Sigmoid function is an S-shaped function. As the independent variable z approaches positive infinity, g(z) approaches 1 infinitely. Conversely, as z approaches negative infinity, g(z) approaches 0 infinitely. It can map any real number to the interval (0, 1), making it suitable for transforming any value function into a function more suitable for binary classification. Due to this property, the Sigmoid function is also regarded as a form of normalization. In linear regression, where $z = \beta^T x$, we can substitute it into the equation, yielding the general form of the binary logistic regression model:

$$g(x) = \frac{1}{1+e^{-x\beta^T}} \tag{4}$$

For binary classification problems, a dependent variable of 1 represents a positive instance, while a dependent variable of 0 represents a negative instance. The aforementioned function has helped us map the actual values to the range of 0 and 1. If g(x) is greater than or equal to 0.5, the predicted output variable is 1. If g(x) is less than 0.5, the predicted output variable is 0.

We can consider g(x) as the probability of the target variable being a positive instance, so 1 - g(x) represents the probability of the target variable being a negative instance. At this point, g(x) takes values between [0, 1]. If we divide g(x) by 1 - g(x) and take the logarithm, we can easily obtain:

$$\ln(\frac{g(x)}{1-g(x)}) = \ln\left(\frac{\frac{1}{1+e^{-x\beta^T}}}{1-\frac{1}{1+e^{-x\beta^T}}}\right) = \ln\left(\frac{1}{e^{-x\beta^T}}\right) = \beta^T x = z \qquad (5)$$

Clearly, the essence of taking the logarithm of odds is z. In other words, it is the logarithm of the odds calculated from the predictions of the linear regression model, which approaches 0 and 1 infinitely. This is why the corresponding model is referred to as “logistic” regression in statistical theory.

Since logistic regression requires the estimation of the parameter β based on the training data and aims to fit the training data as closely as possible, with the goal of achieving a high prediction accuracy on the training set (as close to 100% as possible), a "loss function" is used as an evaluation metric to measure the amount of information loss incurred by the model's fitting to the training set, and to assess the quality of the parameter β. The loss function for logistic regression is derived from maximum likelihood estimation.

Given that our sample output takes two classes, 0 or 1. For a sample i, its class is denoted as $y_i \in (0,1)$. Therefore, for sample i, we can deduce the following relationships:

$$P(\ Y = 1|x_i) = g(x_i) \qquad (6)$$

$$P\ (\ Y = 0|x_i) = 1 - g(x_i) \qquad (7)$$

The likelihood function is expressed as:

$$L(\beta) = \prod[g(x_i)]^{y_i}\,[1 - g(x_i)]^{1-y_i} \qquad (8)$$

For ease of computation, we can take the logarithm of both sides of the equation and simplify, yielding:

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{m} (y_i \log(g(x_i)) + (1 - y_i)\log(1 - g(x_i))) \qquad (9)$$

In machine learning, we have the concept of a loss function, which measures the extent of prediction errors made by the model. In this context, the loss function L (β) is the negative value of the logarithmic likelihood function. The subsequent step involves solving for β that minimizes the loss function. Generally, there are two methods commonly used for logistic regression: gradient descent and Newton's method. Here, we will explain the gradient descent method. The gradient descent method seeks the descent direction by computing the first derivative of the loss function with respect to beta. Through iterative updates, the parameters are updated. Based on the derivation, we can obtain the gradient update formula as follow:

$$\beta_{m+1} = \beta_m - \alpha \frac{\partial L(\beta)}{\partial \beta} \tag{10}$$

Where α represents the learning rate and m is the number of iterations. After each parameter update, the iteration can be stopped when the value of $||L(\beta_{m+1}) - L(\beta_m)||$ is below a threshold or when the maximum number of iterations is reached.

As we strive to minimize the loss function and achieve optimal performance of the model on the training set, it may lead to another problem: if the model performs well on the training set but poorly on the test set, it is said to be overfitting. Although logistic regression and linear regression are inherently underfitting models, it is still necessary to employ techniques to control overfitting and adjust the model. In the case of logistic regression, overfitting is controlled through regularization.

Regardless of how the field of machine learning evolves, logistic regression remains a widely used model due to its irreplaceable advantages:

(1) Logistic regression exhibits excellent fitting performance for linear relationships, especially in cases where there is a strong linear relationship between the features and the labels.

(2) Logistic regression has faster computations under specific conditions. For linear data, logistic regression offers fast fitting and computation, particularly when dealing with large datasets.

4.3 Random Forest

Random Forest has long been regarded as a representative model of the bagging method in the field of machine learning. As the name suggests, it is an inherited algorithm in the machine learning domain, also known as an ensemble learning algorithm. It involves building multiple mutually independent estimators to conduct independent learning. The evaluation results of each estimator's predictions are then averaged to determine the evaluation results of the ensemble estimator, as indicated by this equation:

$$\hat{f}(\mathrm{x}) = \frac{1}{K}\sum_{\mathrm{k=1}}^{\mathrm{K}} f_{\mathrm{k}}(\mathrm{x}) \tag{11}$$

Here, K represents the number of bagging iterations, and $f_{\mathrm{k}}(\mathrm{x})$ denotes each individual base learner.

Alternatively, the evaluation results of the ensemble estimator can be determined using the majority voting principle, as shown in this formula:

$$\hat{f}(\mathrm{x}) = \underset{c\in C}{argmax}\left\{\sum_{\mathrm{k=1}}^{\mathrm{K}} \mathbb{I}(f_k(\mathrm{x}) = \mathrm{c}|\mathrm{c} \in \mathrm{C})\right\} \tag{12}$$

This type of model, derived from multiple individual models, is called an ensemble estimator, where each individual model that constitutes the ensemble estimator is referred to as a base estimator.

Similar to the doctrine of the Mean in Confucianism, ensemble algorithms are not standalone machine learning algorithms themselves. They are a collection of multiple models, where individual models may not exhibit high performance. However, by integrating these individual models, the ensemble often achieves modeling results that surpass those of individual models. This approach is highly valuable in the field of machine learning, as it has been repeatedly observed and summarized that ensemble methods can typically enhance prediction accuracy.

Random Forest utilizes decision trees as the base estimators. During the construction of each tree, only a random subset of features is considered. Typically, the number of features selected is the square root of the total number of features. Unlike conventional bagging algorithms, Random Forest also samples the same number of samples as the training set. The most commonly used base classifier in Random Forest for classification is the Classification and Regression Trees (CART). As a typical supervised learning algorithm, decision trees do not have fixed model parameters and are considered non-parametric. Essentially, decision trees employ inductive learning to discover classification rules from a large set of data samples. In Random Forest, individual decision trees do not require high classification accuracy but should exhibit high diversity among each other. The final classification result is determined by combining the results from all decision trees, which constitutes the core idea of the Random Forest algorithm. The modeling process of Random Forest mainly involves the following two core steps：

(1) Through the bootstrap method, a self-contained sample set is constructed by repeatedly and randomly drawing, with replacement, a number of samples equal to the size of the original training set from the original training set. By repeating the bootstrap sampling process O times, independent bootstrap sample sets are obtained for constructing O decision trees.

(2) Based on the bootstrap sample sets, the CART algorithm is employed to train O decision trees. During the node splitting process of a decision tree, a random subset of attributes is considered. A specific number of attribute subsets are randomly selected from the available attribute set at the current node. The optimal splitting attribute for the current node is determined based on the selected attribute subset and the criterion for attribute selection. Ultimately, a non-pruned classification decision tree is constructed.

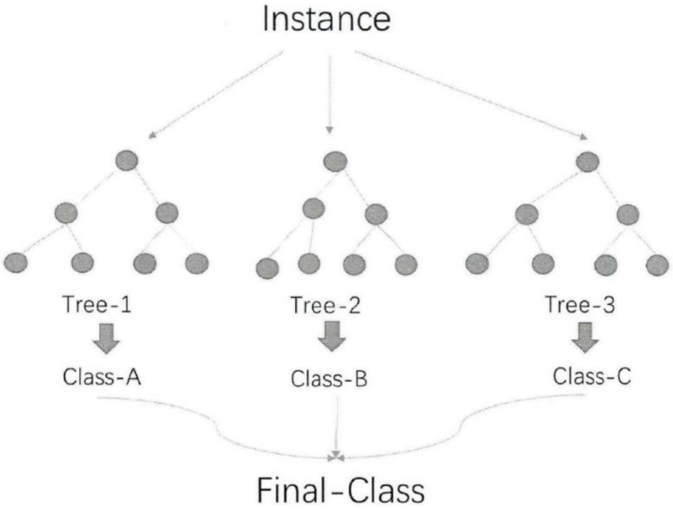

Figure 4. Random forest workflow diagram

The main advantages of the random forest algorithm are as follows: The training of random forests can be highly parallelized, providing advantages in terms of speed when dealing with large sample sets in the era of big data. Random forests are effective in training models even when the dimensionality of the sample features is high, as they utilize random selection of feature splits at tree nodes. The models trained using random forests exhibit low variance and strong generalization ability due to the adoption of random sampling. The introduction of "dual randomness" in the random forest algorithm, achieved through bootstrap sampling and random selection of attribute subsets, helps prevent overfitting and promotes diversity among the base classifiers. During classification, each decision tree participates in a "voting" process to determine the final classification result for a given sample.

The main drawbacks of the random forest algorithm are as follows: When the sample set contains significant noise, the random forest model is still prone to overfitting. The presence of features with a large number of distinct values can have a significant impact on the overall judgment of the random forest.

4.4 Gradient Boosting Decision Tree

Gradient Boosting Decision Trees (GBDT) is a boosting method that combines an additive model with forward distribution algorithm, using decision trees as the ensemble learning model. GBDT is generally trained through multiple iterations, where each iteration generates a weak classifier. These classifiers are designed to target the residuals of the previous classifier, and further improve the training process based on this foundation. Weak classifiers are typically simple and easy to train, often characterized by low variance and high bias. The overall training process aims to continuously reduce the model's bias through iterative steps, with the ultimate goal of maximizing the accuracy of the final model. When using GBDT for classification and regression, the weak learners are typically CART regression trees. Due to the requirements imposed on weak classifiers, the depth of each regression tree is not very deep. The final classifier is obtained by weighting the ensemble of weak classifiers obtained from each iteration, and can be described as follow:

$$\mathrm{F}(\mathrm{x},\ \omega) = \sum_{\mathrm{k}=0}^{\mathrm{K}} \alpha_{\mathrm{k}}\mathrm{h}_{\mathrm{k}}(\mathrm{x},\ \omega_{\mathrm{k}}) \tag{13}$$

The GBDT algorithm can be regarded as an additive model composed of M trees. In the above formula, x represents the input sample, ω represents the model parameters, h represents the classification or regression tree, and α represents the weight of each tree.

The specific algorithmic process for the general loss functions in GBDT is as follows:

(1) Given a training dataset T = $\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$, where $x \in X$, X is the input space, and $L(y, \mathrm{f}(x))$ is the loss function, our objective is to obtain the final regression tree $\mathrm{F_M}$.

(2) The initialization function is denoted as:

$$F_0(x) = \underset{c}{argmin}\sum_{i=1}^{N} L(y_i,c) \tag{14}$$

(3) For i = 1, 2,..., N, calculate the pseudo-residuals for the m-th tree:

$$r_{m,i} = -\left[\frac{\partial L(y_1,F(x_i))}{\partial F(x_i)}\right] F(x) = F_{m-1}(x) \tag{15}$$

(4) For i = 1,2,...,N, fit the data (x_i, $r_{m,i}$) using a CART regression tree to obtain the m-th regression tree, where the corresponding leaf regions are denoted as $R_{m,j}$, with j = 1, 2, ..., J_m, and J_m represents the number of leaf nodes in the m-th regression tree;

(5) For J_m leaf regions, denoted as j = 1, 2, ..., J_m, calculate the optimal fitted values:

$$c_{m,j} = \underset{c}{argmin} \sum_{x_i \in R_{m,j}} L(y_i, F_{m-1}(x_i) + c) \tag{16}$$

(6) Update the strong learner $F_m(x_i)$:

$$F_m(x_i) = F_{m-1}(x_i) + \sum_{j=1}^{J_m} c_{m,j} I(x_i \in R_{m,j}) \tag{17}$$

(6) Let m = m + 1 and repeat steps 1-6.

(7) Continue until the desired number of base learners is achieved, and thus obtain the final model:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J_m} c_{m,j} I(x \in R_{m,j}) \tag{18}$$

The typical approach of using GBDT algorithm to fit regression trees is by utilizing the negative gradient of the loss function. GBDT can be applied to regression or classification tasks. When GBDT is used for binary classification, its loss function is:

$$L(y,f(x)) = \log(1 + \exp(-2yf(x))) \tag{19}$$

Whereas $y \in \{-1,1\}$, $f(x) = \frac{1}{2}\log\left[\frac{\Pr(y=1|x)}{\Pr(y=-1|x)}\right]$, the negative gradient error is $\frac{2y}{1+\exp(2yf_{m-1}(x))}$

Typically, overfitting the training set can lead to a decrease in the generalization ability of the model. To address this, regularization techniques can be applied to the GBDT model by constraining the fitting process to minimize overfitting effects. There are three main methods:

The first method involves multiplying each base model by a weakened coefficient $\mu \in (0,1]$, also known as the learning rate. This reduces the contribution of individual models to the fitting loss, but it also requires a larger number of base learners to build the model. The second method controls the complexity of the base learners by applying regularization pruning techniques. The third method involves training the model on randomly sampled subsets of data, achieved by applying a subsampling ratio. Cross-validation can be used to select the appropriate ratio.

4.5 ROC Curve, AUC Value, f1-score and brier score

The Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) value are commonly used as metrics to assess the goodness of fit of a classification model.

The horizontal axis of the ROC curve is generally referred to as the false positive rate (FPR), and the vertical axis is generally referred to as the true positive rate (TPR). Similarly, there are also true negative rate (TNR) and false negative rate (FNR). The purpose of the ROC curve is to balance the accuracy of identifying positive instances and the error of classifying negative instances as positive. An increase in true positive rate is always accompanied by an increase in false positive rate. The meanings of the four indicators represented by the ROC curve are as follows:

(1) FPR: The probability of falsely classifying a negative instance as a positive instance.

(2) TPR: The probability of correctly classifying a positive instance as a true positive.

(3) FNR: The probability of falsely classifying a true positive as a negative instance.

(4) TNR: The probability of correctly classifying a negative instance as a true negative.

The four values mentioned above are organized into a matrix known as a confusion matrix, as shown in Table 2:

	p	n
p’	TP（True Positive）	FP（False Positive）
n’	FN（False Negative）	TN（True Negative）

Table 3. Confusion Matrix

Where p and n are the true values, p' and n' are the predicted values, we can obtain four metrics: true positive, false positive, false negative, and true negative. Therefore, the following relationships exist:

$$\mathrm{TPR} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \tag{20}$$

$$\mathrm{FPR} = \mathrm{FP} / (\mathrm{FP} + \mathrm{TN}) \tag{21}$$

The AUC value represents the area under the ROC curve, and a larger value indicates better performance of the model. It is worth noting that if the AUC of a model is 0.5, it is no better than random guessing. Therefore, such a model has no predictive value.

F1-score is a machine learning metric based on precision and recall. Precision measures the proportion of true positive predictions among all positive predictions, and is defined as:

$$\mathrm{Precision} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP}) \tag{22}$$

Recall, also known as TPR, measures the proportion of true positive predictions among all actual positive samples, and is defined as:

$$\mathrm{Recall} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \tag{23}$$

The f1-score can be seen as a harmonic mean of precision and recall, serving as a new metric that combines both precision and recall effectively:

$$\text{F1-score} = 2 * \mathrm{Precision} * \mathrm{Recall} / (\mathrm{Precision} + \mathrm{Recall}) \tag{24}$$

The brier score is used to measure the discrepancy between predicted probabilities and true outcomes. It is calculated as the mean squared error of the probability predictions relative to the test samples and is represented as:

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{n} (p_i\text{-}o_i)^2 \tag{25}$$

Here, p_i represents the predicted probability, o_i represents the true outcome, and N represents the sample size. The Brier score ranges from 0 to 1, with a score closer to 0 indicating a smaller discrepancy between the predicted probabilities and the true outcomes, indicating better predictive performance of the model.

5 Results

5.1 Live Stream User Clustering Analysis

Here, we conducted cluster analysis on the data that has undergone missing value analysis, removed customer IDs, but has not yet been normalized. Prior to cluster analysis, we converted all categorical variables into numeric ones. Since each variable may hold certain insights, we performed cluster analysis on all remaining variables.

The elbow method, mentioned earlier, was employed here to determine the value of k. Based on the principle of the elbow method, we calculated the Within-Cluster Sum of Squares (WSS), which is the sum of squared distances between each point and the cluster center. Figure 5 was plotted with the x-axis representing the number of clusters and the y-axis representing the WSS. Typically, such graphs are created to identify the k value at which the squared sum starts to exhibit a bending or flattening pattern, resembling an elbow. In the figure, it is evident that the elbow occurs at k = 4.

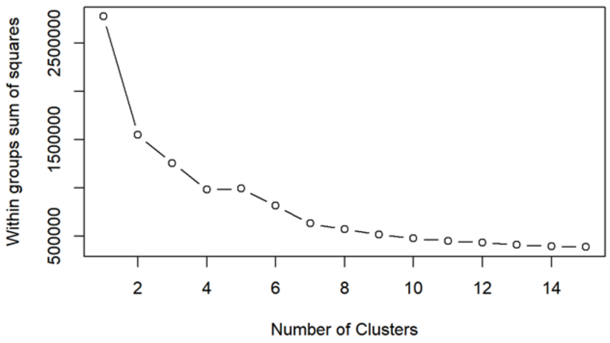

Figure 5. Determining the value of k based on the elbow method

When k=4, we obtained the following four clusters based on the majority of numerical variables. The average values of some numerical variables in each cluster are sorted as follows:

(1) Cluster 1 (14%): Moderate Churn Probability Users. These users have an average distance to the shipping warehouse, with ordinary average monthly app usage time. This cluster had the highest average number of orders in the previous month but the lowest average satisfaction score of orders. They used coupons the most and had the highest average account balance.

(2) Cluster 2 (32%): Low Churn Probability Users. These users have the shortest average distance to the shipping warehouse and the highest average monthly app usage time. This cluster had the highest average satisfaction score of orders.

(3) Cluster 3 (11%): High Churn Probability Users. These users have the longest average distance to the shipping warehouse and the lowest average monthly app usage time. They had the fewest average orders in the previous month. They used coupons the least and had the lowest average account balance.

(4) Cluster 4 (43%): Moderate Churn Probability Users. These users have an average distance to the shipping warehouse, with ordinary average monthly app usage time. They had the highest complaint rate.

Through clustering, it can be observed that the average distance between users and the shipping warehouse is positively correlated with the likelihood of churn, as is the average monthly app usage time. Customers with lower account balances are also more prone to churn. These findings align with our understanding of e-commerce in practice. Typically, we would assume that customers with lower satisfaction levels are most likely to churn. However, the above definition reveals that this may not always be the case. Although Cluster 2 exhibits the highest average satisfaction score and represents customers with a lower likelihood of churn, having the lowest average satisfaction score does not necessarily imply the highest likelihood of churn.

5.2 Variable correlation analysis

Due to the research objective of making targeted changes to marketing strategies, it is necessary to select relevant features that are aligned with the business. In other words, we can use these features to design new marketing campaigns. Initially, we explore the correlation between variables by using a variable correlation matrix plot, as shown in Figure 6. It can be observed that there are no dots in the first row for “Ordercount” and “CouponUsed”. This indicates that the p-values are greater than 0.05. Based on this, these two variables will be excluded in subsequent analyses.

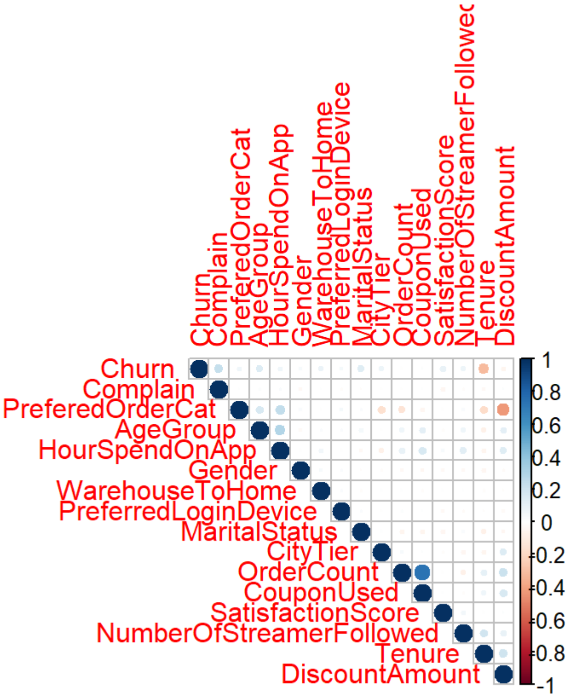

Figure 6. Variable correlation matrix plot

5.3 Feature Importance Analysis

Based on the clustering analysis, feature selection, and data preprocessing mentioned earlier, we conducted regression analysis on the four identified clusters using the “glm” function in R. In this analysis, we focused only on Cluster 2 and Cluster 3, as they represent the low-churn and high-churn clusters, respectively. The results and their practical interpretations are presented as follows. It is evident that the duration of platform usage and whether there were complaints in the previous month are the key factors:

(1) Cluster 2, low-churn customers: These users not only exhibit sensitivity towards platform

usage time and complaints but also pay attention to the distance between the delivery warehouse and themselves. As shorter distances result in faster delivery, we can infer that they are concerned about delivery duration. If the customer resides in a third-tier city, they are more likely to churn. The number of anchors followed by individual users may appear unrelated to the research question initially, but it can actually provide valuable insights. An account that follows a larger number of anchors indicates a broader range of interests, suggesting that the user may not be a devoted fan of our specific live-streaming room. Conversely, a smaller number of followed anchors indicates that the user is more inclined to watch streams from a few specific live-streaming rooms. A loyal fan of a brand or live-streaming room is less likely to be prone to churn, as they are less likely to easily switch to other platforms or content providers.

Coefficients	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-4.5126	0.7054	-6.397	1.58e-10	***
Tenure	-1.8804	0.2662	-7.063	1.63e-12	***
PreferredLoginDevice	0.3063	0.1491	2.054	0.03996	*
CityTier	0.4438	0.1482	2.994	0.00275	**
WarehouseToHome	0.3700	0.1387	2.668	0.00764	**
MaritalaStatus	0.3545	0.1474	2.406	0.01614	*
Gender	0.2150	0.1521	1.413	0.15759	
PreferedOrderCat	-0.3167	0.1432	-2.212	0.02696	*
NumberOfStreamerFollowed	0.5992	0.1499	3.998	6.39e-05	***
Complain	0.5373	0.1366	3.934	8.35e-05	***
DiscountAmount	1.2478	0.5328	2.342	0.01917	*

Table 4. The logistic regression analysis results for cluster 2

For Cluster 2, we employed the random forest algorithm to obtain a variable importance plot. The purpose was to assess the relative importance of variables in predicting customer churn within this specific cluster. The results were generally consistent with the findings from the logistic regression analysis, with the difference being that random forest identified higher importance for variables such as "DiscountAmount" and "MaritalStatus".

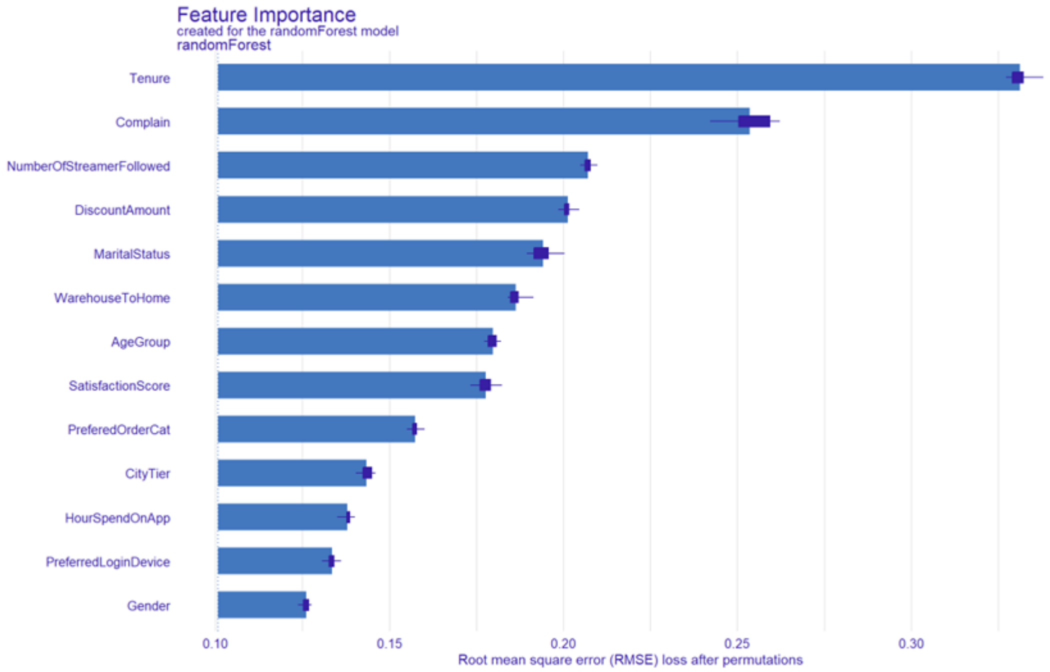

Figure 7. The feature importance plot for Cluster 2

(2) Cluster 3, high-churn customers: The churn of customers within this cluster is influenced by various factors, among which Tenure, WarehouseToHome, NumberOfStreamerFollowed, and Complain are the most significant. According to Table 5, the longer the duration of using the live streaming platform, the less likely they are to churn. Moreover, customers are more prone to churn if they follow a greater number of streamers, reside farther from the warehouse, or have a history of complaints. Interestingly, as age increases, customers in Cluster 3 are more likely to churn.

Coefficients	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-3.0571	0.2083	-14.676	<2e-16	***
Tenure	-1.9166	0.2363	-8.112	4.98e-16	***
PreferredLoginDevice	0.1824	0.1230	1.483	0.138058	
CityTier	0.1815	0.1183	1.534	0.125012	
WarehouseToHome	0.3893	0.1116	3.4893.313	0.000495	***
AgeGroup	0.3278	0.1379	2.377	0.017435	*
PreferedorderCat	-0.4500	0.1456	-3.091	0.001994	**
NumberOfStreamerFollowed	0.6003	0.1187	5.056	4.27e-07	***
SatisfactionScore	0.2935	0.1257	2.335	0.019561	*
Complain	0.7639	0.1194	6.395	1.60e-10	***

Table 5. The logistic regression analysis results for cluster 3

In this context, for Cluster 3 - high probability churn customers, we also applied the Random Forest algorithm to generate a variable importance plot.

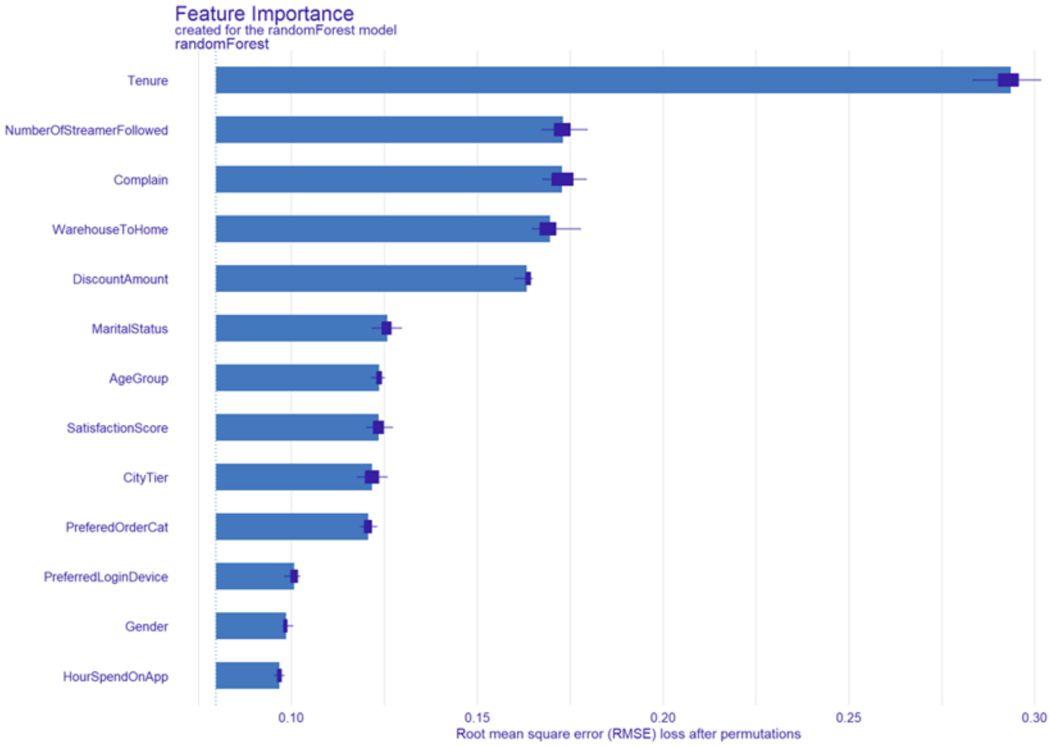

Figure 8. The feature importance plot for Cluster 3

We can observe that, in addition to the four significant features identified in logistic regression, the feature depicting the average cashback amount from the previous month in the plot is also relatively important in predicting customer churn.

5.4 Churn Analysis

The aim of this study is to predict whether users will continue placing orders in the same live streaming room in the second month, indicating customer churn. As it is a classification problem, supervised models are employed for prediction. In the previous section's feature analysis, logistic regression was utilized for user prediction; however, the algorithm's accuracy was found to be relatively low during the model evaluation process. After reviewing relevant literature and resources, it was discovered that random forest models and Gradient Boosting Decision Tree (GBDT) often yield better prediction results in binary classification models. Therefore, this study also attempted to compare and analyze these two models with logistic regression, with the hope of improving the predictive performance of the models. Furthermore, in terms of data preparation, a 30% random sample was extracted from each cluster as the testing set, while the remaining data was used as the training set. Moreover, as mentioned earlier, the data types of various indicator variables have been adjusted in the previous steps, allowing for direct execution of the relevant operations to initiate the modeling process. Similarly, here we only focus on Cluster 2 and Cluster 3.

5.4.1 Prediction and Model Evaluation

Based on the confusion matrix, it can be observed that for Cluster 2 and 3, although the logistic regression model aids in understanding feature importance, its predictive capability for churn detection is slightly inferior to tree-based models. Based on various metrics, both Random Forest and GBDT outperform the logistic regression model, regardless of whether it is Cluster 2 or Cluster 3.

Logistic Regression		**Random Forest**		**GBDT**	
Cluster 2	Value	Cluster 2	Value	Cluster 2	Value
Accuracy	90.91%	Accuracy	92.88%	Accuracy	92.88%
Precision	91.28%	Precision	92.69%	Precision	93.27%
Recall	96.37%	Recall	99.69%	Recall	99.03%
F1-score	0.9376	F1-score	0.9607	F1-score	0.9606
Cluster 3	Value	Cluster 3	Value	Cluster 3	Value
Accuracy	87.8%	Accuracy	92.04%	Accuracy	94.45%
Precision	93.68%	Precision	96.08%	Precision	95.87%
Recall	90.82%	Recall	98.99%	Recall	97.89%
F1-score	0.9232	F1-score	0.9751	F1-score	0.9687

Table 6. Comparison of Model Confusion Matrices

Evidently, whether in Cluster 2 or Cluster 3, all three models have demonstrated respectable AUC values. The model performance of Random Forest and Gradient Boosting Decision Trees (GBDT) is closely comparable, and both outperform Logistic Regression.

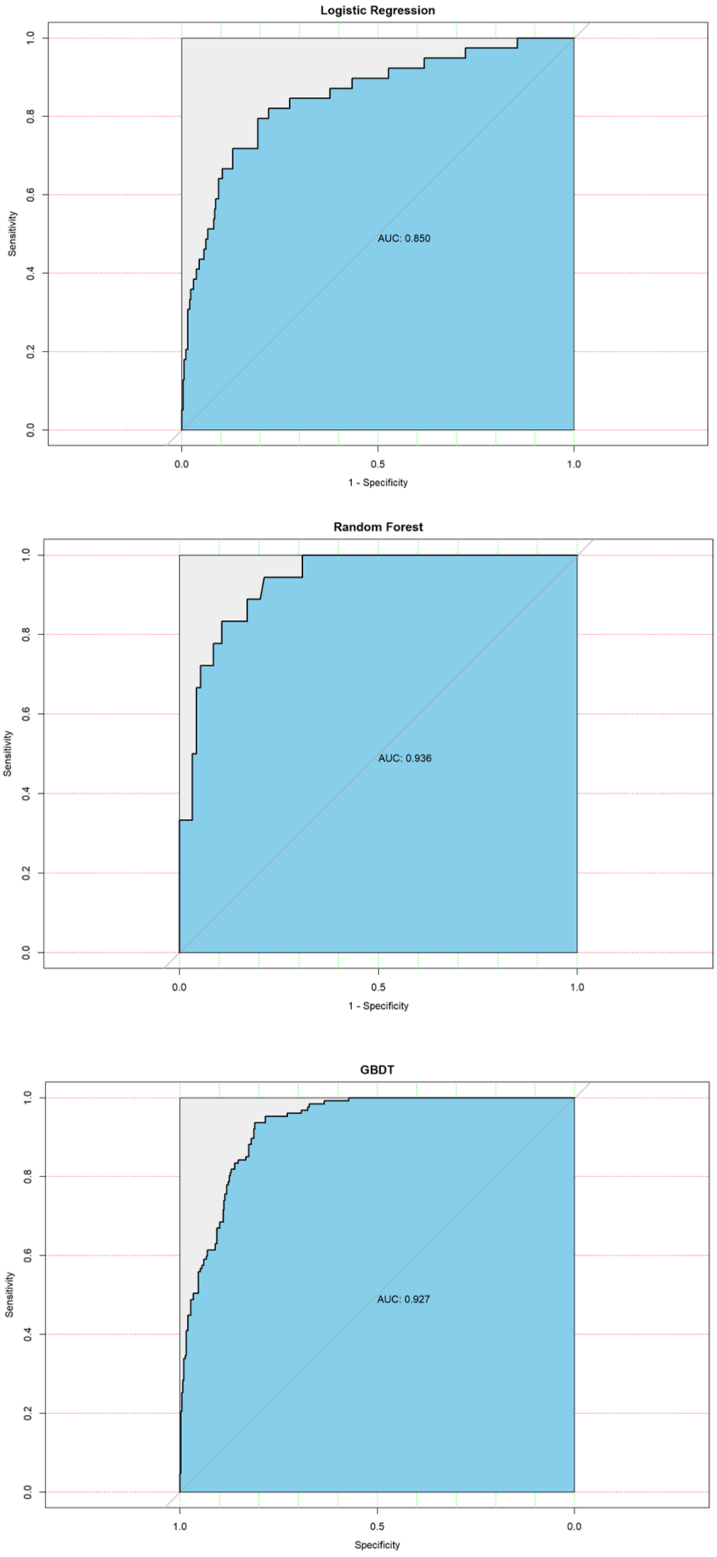

Figure 9. Comparison of ROC curves for cluster 2

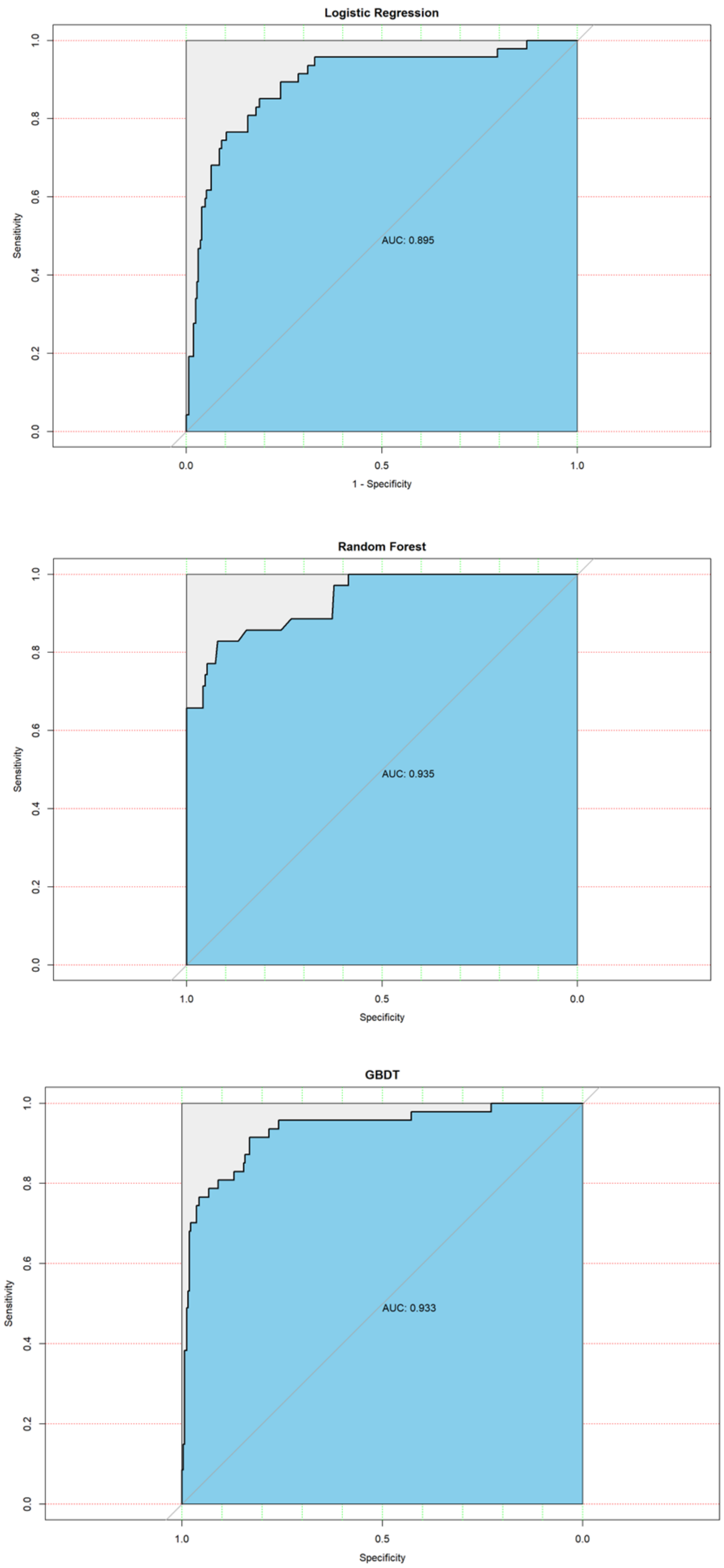

Figure 10. Comparison of ROC curves for cluster 3

Temporarily excluding logistic regression, for Cluster 2, the Brier scores for Random Forest and GBDT are 0.132 and 0.163, respectively. For Cluster 3, the Brier scores for Random Forest and GBDT are 0.134 and 0.151, respectively. Therefore, based on the magnitude of the Brier scores, Random Forest demonstrates superior performance.

Both Random Forest and GBDT have made it possible to accurately identify churned customers. In the future, we can continue to explore the addition of more user-specific features, as the backend data of live streaming platforms is expected to become more abundant and detailed. For instance, future backend data collection could include the duration of a user's stay in a particular live streaming session. At the very least, based on the existing dimensions, Random Forest and GBDT have already met our requirements for accurately identifying customers with a high probability of churn.

6 Conclusion, Recommendations, and Outlook

6.1 Conclusion and Recommendations

From the cluster analysis, it can be observed that the distance between warehouses and customers is highly correlated with cluster classification, where clusters with greater average distance tend to have higher churn rates. This indicates that larger distance values are associated with a higher likelihood of customer churn. When considering the practical implications, this relationship is not coincidental as greater distance implies longer delivery times. The waiting time for customers' previous purchases, which is influenced by delivery time, can contribute to their churn. Improving this aspect requires more focus on logistics research. However, the significance of cluster analysis becomes evident in identifying the factors contributing to customer churn. The usage duration of the live streaming platform in the previous month is also highly correlated with cluster classification, with shorter usage duration indicating a higher likelihood of customer churn. The live streaming platform can explore the possibility of introducing a live-streaming-exclusive promotional campaign where viewers can redeem coupons based on their viewing duration, to prolong customers' engagement. Although this may alter the nature of the variable itself, collecting data and conducting a similar analysis after implementing this approach can validate its effectiveness. As is well known, customer after-sales service and follow-up have always been a matter of concern for various sellers. As expected, customer satisfaction and complaints are related to customer churn. Improving customer satisfaction and providing dedicated after-sales service are certain to alleviate customer churn.

The core indicators shared by Cluster 2 and Cluster 3 are the customer's usage duration on the live platform in February and whether they filed a complaint during that month. These two factors are crucial for customer retention regardless of the user category. For customers with a lower likelihood of churn, those who purchase main products are less likely to churn. This can be easily understood as customers who buy laptops and smartphones are more likely to require additional accessories in the future. Therefore, intensifying promotions for laptops and smartphones may attract a larger customer base and provide more potential buyers for related accessories. Based on the core indicators of Cluster 3, we can observe that the customer's age do impact the churn probability. Generally, different demographics have varying preferences for live content. In light of this, the live streaming operation team needs to strike a balance between catering to the existing customer base and making changes that may lead to the churn of current customers in order to accommodate the preferences of different demographic groups. Engaging in targeted marketing for specific customer segments at appropriate opportunities could be a viable approach without compromising customer privacy. For example, launching a promotion targeting a specific age group of males during Father's Day. By ensuring the confidentiality of customer personal information, such methods might be able to leverage high-churn probability consumer groups that have been persistently difficult to retain.

The analysis of feature importance after clustering reveals that the core behaviors and attributes of different customer types are not consistent. In addition to adjusting the models using recommendation algorithms, it is recommended to employ clustering techniques to guide different customer behaviors and attract them to engage in core actions or induce them to become customers with a high probability of retention. This approach helps the operation team gain a better understanding of the live streaming platform and its customers, thereby enhancing customer experience and improving the effectiveness of marketing efforts.

Considering all evaluation criteria, among the models trained in this study, Random Forest is the most suitable for predicting customer churn in the given context. In practical application, it is advisable to implement targeted customer retention strategies based on the algorithm's predictions of churned customers, such as delivering discount coupons through email campaigns specifically targeted at customers identified as at risk of churn.

In summary, the model predictions help us identify potential churned customers, while clustering analysis helps us understand the underlying reasons for potential churn. Combining both approaches enables us to develop targeted marketing strategies. The overall process can be as follows. Collecting data at the end of the month, using the model to predict churned customers, and then designing targeted marketing strategies based on clustering analysis. Data can be collected again at the end of the following month for evaluation and reflection. In practical applications, the live streaming platform can establish this iterative process of analysis and adaptation.

6.2 Limitations and Future Directions of the Study

Current Internet e-commerce competition is exceptionally fierce. However, the format of live-streaming e-commerce is not yet popular in most countries or regions. For a relatively new form of e-commerce, a simplistic approach to customer operations and a single-target formula in algorithm models may not meet customer demands adequately, as each country or region has its own characteristics. This is especially true for live-streaming platforms that offer a wide variety of products, where different personalized ranking models after customer segmentation can be explored as future directions.

In real customer data, data imbalance is a common and unavoidable issue, particularly in churn prediction models. Confronting imbalanced training datasets, many models may yield unsatisfactory results, warranting further in-depth exploration and research.

This paper employed three classification algorithms, namely logistic regression, random forest, and GBDT, to predict customer churn using empirical data. Although the performance was promising, in the current booming era of big data, the data resources from various customer dimensions are becoming increasingly vast and diverse. The construction of predictive models for relevant live-streaming platforms requires the analysis of a wider range of indicators and dimensions. However, due to limitations such as data sensitivity, computational and statistical techniques, this study only analyzed a few customer attributes mentioned prominently. In the future, there is ample room for exploration and the necessity for further in-depth research in this regard.

Furthermore, even if the estimation template used is consistent, once applied to different products, different live-streaming platforms, and different regions, it may yield different results in the analysis. Therefore, independent thinking and research are required to construct tailored estimation approaches based on data feedback in order to predict user churn and enhance the efficiency of addressing practical issues.

To date, with the rise of content platforms, especially short video platforms, there has been an increasing influx of users to e-commerce live-streaming, resulting in a growing number of choices for e-commerce live-streaming customers. This has led to escalating customer churn in various live-streaming platforms and rooms. However, there is still a lack of research in the industry and academic community regarding customer attrition in e-commerce live-streaming. At the same time, a significant amount of customer data is generated during the regular operation of live-streaming rooms, and the data is updated rapidly. Thus, in order to ensure the practicality of the customer churn prevention system, promptly responding to the operational situation of live-streaming rooms and enhancing the targeted nature of marketing, it is necessary to conduct research on customer churn prediction in live-streaming rooms and establish a more robust operational mechanism to ensure the long-term healthy operation of live-streaming rooms.

References

China Internet Network Information Center. (2023). China Internet DevelopmentStatus Statistical Report. Retrieved April 3, 2023, from *https://www.cnnic.net.cn/n4/2023/0303/c88-10757.html*.

Wongkitrungrueng, A., & Assarut, N. (2020). The role of live streaming in building consumer trust and engagement with social commerce sellers. *Journal of Business Research*, *117*, 543-556.

Kumarab, V., & Venkatesanc, R. (2021). Transformation of metrics and analytics in retailing: The way forward. *Journal of Retailing, 97*, 496-506.

Liu, Z. (2020). Research on the current situation and future trend of web celebrity e-commerce live streaming industry. *The 4th International Conference on Business and Information Management*.

Cunningham, S., Craig, D., & Lv, J. (2019). China's livestreaming industry: Platforms, politics, and precarity. *International Journal of Cultural Studies*, *22*(6), 719-736.

Sun, Y., & Li, X. (2019). How live streaming influences purchase intentions in social commerce: An IT affordance perspective. *Electronic Commerce Research and Applications*, 37.

Wang, D., Luo, X. R., Hua, Y., & Benitez, J. (2022). Big arena, small potatoes: a mixed-methods investigation of atmospheric cues in live-streaming e-commerce. *Decision Support Systems*(Jul.), *158*.

Cui, X.Q., Li, Y.J., Li, X., & Fang, S.L. (2023). Livestream e-commerce in a platform supply chain: A product-fit uncertainty reduction perspective. *International Journal of Production Economics*, *258*.

Chang, Y.T., Yu, H., & Lu, H.P. (2015). Persuasive messages, popularity cohesion, and message diffusion in social media marketing. *Journal of Business Research*, *68*(4), 777-782.

Chen, Z., Benbasat, I., & Cenfetelli, R.T. (2017). "Grassroots internet celebrity plus live streaming" activating IT-Mediated lifestyle marketing services at e-commerce websites. *International Conference on Interaction Sciences*.

Geng, R., Wang, S., Chen, X., Song, D., & Yu, J. (2020). Content marketing in e-commerce platforms in the internet celebrity economy. *Industrial Management & Data Systems*, *120*(3), 464-485.

Meng, L. M., Duan, S., Zhao, Y., Kevin, L., & Chen, S. (2021). The impact of online celebrity in livestreaming e-commerce on purchase intention from the perspective of emotional contagion. *Journal of Retailing and Consumer Services*, *63*.

Zhu, L.J., & Li, H.Y. (2021). How do Anchors' characteristics influence consumers' behavioral intention in livestream shopping? A moderated chain-mediation explanatory model. *Frontiers in Psychology*, *12*.

Ha, I., Yoon, Y., & Choi, M. (2007). Determinants of adoption of mobile games under mobile broadband wireless access environment. *Information and Management*, *44*(3).

Ling, Z., Lu, Y., & Wang, B. (2011). What makes them happy and curious online? An empirical study on high school students' Internet use from a self-determination theory perspective. *Computers and Education*, *56*(2), 346-356.

Wang, X., & Wu, D. (2019). Understanding user engagement mechanisms on a live streaming platform. *Orlando: HCI in Business*, 266-275.

Lu, W. (2019). Research on e-commerce mode based on live delivery. *2019 7th International Education, Economics, Social Science, Arts, Sports and Management Engineering Conference*.

Bickart, B. & Schindler, R.M. (2001). Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*, *15*(3).

Singh, J.P., Irani, S., & Rana, N.P. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, *70*.

Hu, M., Zhang, M., & Wang, Y. (2017). Why do audiences choose to keep watching on live video streaming platforms? An explanation of dual identification framework. *Computers in Human Behavior*, *75*, 594-606.

Steward, S. (1996). Technology springs forward to melt churn. *Cellular Business*, *13*(5), 30-34.

Sladojevic, S.M., Culibrk, D.R., & Crnojevic, V.S. (2011). Predicting the churn of telecommunication service users using open source data mining tools. *International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services*, 749-752.

Gursoy, U.T.S. (2010). Customer churn analysis in telecommunication sector. *Istanbul University Journal of the School of Business*. *39*(1), 35-59.

Kim, KH., Lee, CS., Jo, SM., & Cho, SB. (2015). Predicting the success of bank telemarketing using deep convolutional neural network. *International Conference on Soft Computing and Pattern Recognition*, 314-317.

Long, X., Yin, W., & An, L. (2012). Churn analysis of online social network users using data mining techniques. *International Multiconference of Engineers and Computer Scientists*.

Liu, D.S., & Ju, C.H. (2009). Customer churn analysis model in manufacturing industry. *Ultra-precision Machining Technologies*.

Patrick, L.S. (2001). Designing interactive value development: perspective and strategies for high precision marketing. *Lund Studies in Economics & Management*, 360.

Chaudry, A.S., Azmat, S., & Sohail, M. (2018). State contingent and conventional banking: The optimal banking choice model. *Economic Modelling*, *65*, 167-177.

Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, *7*(1), 9.

Yang, S.H., & Zhang H.M. (2018). Comparison of several data mining methods in credit card default prediction. *Intelligent Information Management*, *10*, 6-14.

Claudio, A., Vincenzo, F., & Ruggiero, S. (2011). Data mining in real estate appraisal: A model tree and multivariate adaptive regression spline approach. *Aestimum*, *58*, 27-45.

Wang, L., Xu, H., & Cao, Y. (2018). Research and implementation of precision marketing system based on big data analysis. *Iop conference*, 394.