Journal of Global Information Management
Volume 30 • Issue 3

Improving Customer Value Index and Consumption Forecasts Using a Weighted RFM Model and Machine Learning Algorithms

Zongxiao Wu, Southwestern University of Finance and Economics, Chengdu, China

Cong Zang, Southeast University, Nanjing, China

Chia-Huei Wu, Institute of Service Industries and Management, Minghsin University of Science Technology, Hsinchu, Taiwan

https://orcid.org/0000-0001-6399-2113

Zilin Deng, Southwestern University of Finance and Economics, Chengdu, China

Xuefeng Shao, Newcastle Business School, The University of Newcastle, Callaghan, Australia

Wei Liu, Business School, Qingdao University, Qingdao, China

https://orcid.org/0000-0002-3044-9795

ABSTRACT

Collecting and mining customer consumption data is crucial to assess customer value and predict customer consumption behaviors. This paper proposes a new procedure, based on an improved random forest model, by adding a new indicator joining the RFMS-based method to a K-means algorithm with the entropy weight method applied in computing the customer value index, classifying customers to different categories, and then constructing a consumption forecasting model whose RMSE is the smallest in all kinds of data mining models. The results show that identifying customers by this improved RMF model and customer value index facilitates customer profiling, and forecasting customer consumption enables the development of more precise marketing strategies.

KEYWORDS

Computing, Consumer, Consumption Forecast, Data Mining, K-Means Clustering Analysis, Marketing Strategy, Random Forest Model, RFM Model

INTRODUCTION

The game industry spans ages and countries and is expanding through game-related platforms such as e-sports competitions and live games. In the context of the outstanding performance of key game products of head game companies; the gradual process of new version approval release; the rapid development of cloud games after the commercialization of 5G as well as the development of VR/AR games - quickly gaining a player's favor with an attractive marketing strategy is very important in this highly competitive industry.

DOI: 10.4018/JGIM.20220701.oa1

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Gartner (2014) defined big data as a massive, high-growth, and diverse information asset that requires new processing models to have greater decision making, insight and process optimization capabilities. The famous futurist Toffler (1980) portrayed "big data" as "the third wave of cadenza" in his book "The Third Wave". Today's big data development is expanding globally with apparently unlimited business opportunities. Having recognized the influence of big data on various industries the major developed economies, such as the United States and European Union, are actively promoting big data strategies. China also regards it as an important support for their new economy. Data mining generally refers to processes of searching for more specific information, hidden in large amounts of general data. Mining is usually based on a database, composed of large amounts of accumulated data, and the process draws out potentially valuable information to support or improve decision making.

Data mining is a powerful technique to help companies identify patterns and trends in their customers' data, and then drive improved customer relationships; accordingly, it is a well-known tool in customer relationship management (CRM). Based on traditional methods of data mining, Cheng and Chen (2008) proposed a new procedure, joining quantitative value of RFM attributes and K-means algorithms into rough set theory (RS theory), to extract meaning rules. This development can effectively reduce drawbacks in previous data mining tools. At the same time, data mining has a tremendous advantage for researchers, because it enables them to extract further hidden knowledge that has been inherited in the raw data. The study of Ravasan and Mansouri (2015) explained a brand new and practical fuzzy analytic network process (FANP), based on a weighted RFM (Recency, Frequency, Monetary value) model for application in a K-means algorithm for auto insurance customers' segmentation. Previously, Hsieh (2004) had proposed an integrated data mining and behavioral scoring model, to manage existing credit card customers more effectively in a bank.

As a new field application, game data mining is emerging, with customer segmentation and customer feature mining the most important entry points. Achieving more precise customer segmentation and more effective customer feature extraction is a long-standing technical difficulty in the game industry (Shahri et al., 2019). Clarifying the customer value measurement index, constructing a customer value identification path, classifying customer value, as well as mining the core consumer and key consumer group characteristics are all necessary for better operation. This paper takes the stored value data of a certain game of a famous Chinese game company as the research sample, to undertake the following investigations: use the improved RFM model to score each player's value index, and classify the players according to the indicators in this model; utilize logistic regression, decision tree, SVM and other models to predict the total amount of stored value of each new player in the future, and evaluate the accuracy of different models; select the best model and optimize it to predict the amount of stored value of players; finally, according to the players' distribution, heterogeneity and stored value behavior, to provide more reliable reference indicators for corporate marketing and customer relationship management. These studies are designed to ensure that companies and marketers can better understand their customers, and then choose differentiated marketing tools to assist with the business development. From the perspective of market space and policy trends, the overall solution of game operators' customer segmentation and prediction of future benefits from new customers, can be used as a benchmark that can positively impact the game industry and act as an indicator for other industries.

LITERATURE REVIEW

Customer Value Forecasting

The study of customer value prediction began in the 1990s (Woodruff, 1997) and there are now many classifications used in establishing models, such as regression analysis, decision tree and naive Bayes. But recently, random forests, artificial neural networks, SVM, Bayesian networks and other algorithms are more frequently used to predict the behaviors of customers. Jahromi (2009) established

a two-step model for telecom data. First, the customers were clustered according to RFM data into four categories, and then neural network and different decision tree methods were applied to these four types of customers, for prediction. Each type of data uses different classification methods, and the predictions were found to be more accurate. The empirical evidence shows that Classification and Regression Tree (CART) is better than C5.0, but the C5.0 method is more accurate and effective than CART in the success rate of a new data set. The K-means algorithm was used to cluster the customers according to their characteristics, and then the CART decision tree algorithm was used for prediction. Kasiran, Ibrahim and Ribuan (2014) proposed two recurrent neural network models for data from the telecommunications industry and used these two models to establish a customer prediction model. The results show that the Jordan Recursive Neural Network model gives higher prediction accuracy than the Elman Recurrent Neural Network model.

RFM Model

The RFM model is an important tool for measuring customer value and customer profitability; it is widely referred to in customer relationship management literature. The model describes the customer's value by his recent purchase behavior, the overall frequency of purchases and how much money is spent. Many studies have used both RFM models and actual industry strategies.

The study by Coussement, Benoit and Van den Poel (2010) showed that a decision tree is better than RFM analysis and logistic regression, when data are of high accuracy, and recommended the use of a decision tree for direct marketing in the context of customer segmentation. Khajvand et al. (2011) used customer lifetime value (CLV) for segmentation in health and beauty companies. Initially, the RFM marketing analysis method is used for customer segmentation. Then, the proposed extended RFM analysis method has an additional parameter. However, after comparing the results, it was found that the addition of new parameter had no effect on the clustering results. Khajvand and Tarokh (2011) derived the customer's future value based on the segmentation of customer lifetime value. They then estimated the future value of the customer based on adaptive weighted RFM analysis, which is applicable to each customer segment within the retail banking range. Khobzi, Akhondzadeh-Noughabi and Minaei-Bidgoli (2014) added two variables, the average frequency and monetary guild, into the RFM model for customer life value analysis. Together with the basic classification model, the classification of online payment customers of the bank was studied. Dursun and Caber (2016) used data mining technology to conduct RFM analysis of hotel customers, pointing out that managers should propose new management strategies based on changes in RFM indicators, when implementing customer relationship management.

Random Forest Algorithm

The Random Forest Classification algorithm is a relatively new algorithm that has performed well and is widely used in many fields. Beckschafer et al. (2014) proposed a subtropical leaf area index calculation method, based on random forests. Hobley, Baldock and Wilson (2016) used random forests to study the effects of environment and human activities on organic carbon components in soils. Gounaridis and Koukoulas (2016) analyzed the main variables affecting soil type. Behrens, Pierdzioch and Risse (2018) studied the joint efficiency of forecasts by means of multivariate random forests, which they used to model the links between forecast errors and predictor variables in a forecaster's information set.

Tanizaki et al. (2019) proposed demand forecasting in restaurants using machine learning. Lohrmann and Luukka (2019) interpreted the classification of the S&P500 open-to-close returns as a four-class problem and compared four trading strategies based on a random forest classifier to a buy-and-hold strategy. Herrera et al. (2019) compared the long horizon forecast performance of traditional econometric models with machine learning methods (Neural Networks and Random Forests), for the main energy commodities in the world, using monthly prices provided by the International Monetary Fund (IMF). Ciner (2019) showed that when the random forest method, which accounts for both linear

and nonlinear dynamics, was used for regression, then industry returns indeed contain significant out of sample forecasting power for the market index return. Behrens and Pierdzioch (2018) documented that optimality of longer-term inflation forecasts, using random forests, cannot be rejected and that forecasts were weakly related to estimates of four German research institutes. Gao and Lu (2015) introduced a novel non-linear regression method: random forest regression (RFR), to quantitatively estimate China railway freight volume. Feng and Wang (2017) proposed a random forest model and a GBM packet - a method to increase the capacity of random decision tree – to improve the decision tree. The results and accuracy of multiple regression analysis are greatly improved when the random forest model is used to predict the demand for bicycles. Chen et al. (2012) proposed a stepped model that first organizes random forest (RF) and core of rough set exploration systems (RSES) to construct various combinations of extracted key attributes for reducing data dimensions; they then developed an indicator that represents the financial status and operational competence of Asian banks for parties interested in investing in Asia. Lessmann (2017) found that random forest regression was particularly effective for resale price prediction.

METHODS

Game Industry Terminology

The game active user group isrelated to "game churning users", which refers to users who will log into the game from time to time and bring consumer value to the game company. Active game users are widely used in the development of the game industry to measure the current state of the game. The main terms that appear in the analysis statistics of online services, such as online documents, webmail services, online games and SNS games are generally defined below (Fabisiak, 2018).

- **Daily Active Use (DAU):** A method or standard term which is used to measure liveliness in online games, social networking services, and mobile apps on a daily timeline; it can also be used to determine the success rate of these services.
- **Monthly Active Users (MAU):** A method or standard term which is used to measure liveliness in online games, social networking services, and mobile apps on a monthly timeline. It can also be used to determine the success rate of these services. Typically, this metric is measured by counting the number of unique users during a particular measurement period (eg, within the previous 30 days).
- **Stickiness of Users (SoU):** A static description of a game user's activity on a monthly timeline; it represents the user's frequency of visits per unit time, the higher the value, the higher the user's stickiness of the game.

  SoU = DAU/MAU

- **Average Revenue Per User (ARPU):** An average profit that a game operator gets from each user on a monthly timeline. The more high-class users, the higher the ARPU, the higher the company's current profit value, the better development prospects and investment feasibility.

  ARPU = Monthly Revenue/Monthly Paying Users

- **Average Revenue Per Paying User (ARPPU):** An average profit that a game operator receives from each paying user on a monthly timeline. This reflects the average amount of payment per paying user.

ARPPU = Total Cycle Payment/Cycle Payments

Figure 1. Customer categories divided by RFMS indicators

R ↑ F ↑ M ↑ S ↑ important value	R ↑ F ↓ M ↓ S ↑ general value
R ↑ F ↑ M ↑ S ↓ important maintenance	R ↑ F ↓ M ↓ S ↓ general maintenance
R ↑ F ↑ M ↓ S ↑ important development	R ↓ F ↓ M ↓ S ↑ general development
R ↑ F ↓ M ↑ S ↓ important retention	R ↓ F ↓ M ↓ S ↓ general retention

Player Classifications in the Improved RFM Model

As an exploratory analysis method, the RFM model exploits the economic value of players through their consumption behaviors and uses these statistics to promote business development and customer relationship management. The following are specific indicators of the RFMS model, which ranks customers based on the four indicators, assuming that the scores of the four indicators are from 1 to 5, respectively.

- **R (Last consumption):** Customers who have recently purchased products or services are most likely to be consumers who are coming back and will be the first to respond to the latest products. In this improved RFM model, the R value is calculated based on the player's last stored value date. The closer the stored value date is to the current date, the higher the R score.
- **F (Total consumption frequency):** Customers who purchase the products and services with the highest frequency, within a limited time, are also the customers with the highest satisfaction about this game. In this improved model, the F value is calculated based on the player's stored value frequency. The higher the stored value frequency, the higher the F score.
- **M (Total amount of consumption):** In a limited time, according to Pareto's Law, 80% of profits come from 20% of important customers. The higher the total amount a customer spends, the more the company needs to maintain the relationship. In this model, the M value is based on the total amount of the stored value of the player. The higher the total value of the stored value in the time period, the higher the score.

This study adds an additional indicator S to improve the traditional RFM model - to the RFMS model. S represents the standard deviation of the customer's stored value recent period of time. It measures the fluctuation of the customer's consumption behavior. The larger the standard deviation, the greater the fluctuation of the customer's stored value each time, which represents impulsive consumption but not sticky consumption.

After calculating the four indicators' scores in the RFMS model, the number of customers in each category is compared according to the average of R, F, M and S. The standard division refers to the difference above or below the average value of each index. The customers are separated into the 8 categories summarized in Figure 1.

Determining Index Weights using the Entropy Weight Method

Entropy was first introduced by Shannon (1948) as a standard to measure disorder. In a comprehensive assessment, evaluating each index is a complex decision-making process of multi-attributes, multi-levels and multi-objectives; the widely used entropy weight method is an objective valuation method. In this paper, the entropy weight method is used to measure the weight of the four pointers of the RFMS to calculate the total value index, so that the weight obtained is much more precise than the subjective definition. The mark is used to indicate the score of a player in each indicator (1-5 points). The total score after considering each indicator, represents the player value index. It is a comprehensive RFMS

model index and if the value is higher than the average, the game company will consider developing further markets for these high-value players.

Player Grouping and Feature Analysis: K-Means Cluster Analysis

The K-Means algorithm is a typical distance-based clustering algorithm. It uses distance as the evaluation index of similarity - the closer the two objects, the greater their similarity. The principle is to first select K objects from a set of n as the initial cluster center; for the remaining objects, they are assigned to the similar category (represented by the cluster center) according to their similarity (distance) with these cluster centers. Then the cluster centers of each new cluster are calculated and the process is repeated until the standard measure function is reached, which is generally when the mean square error function begins to converge. The unique feature of this method is that the distance between each cluster group is the smallest and the distance between groups is the largest.

Reduction of High-Dimensional Data Using the TSNE Algorithm

The goal of TSNE is to aggregate small "neighbors" between similar data points, while reducing the dimension of the overall data to make it easier to visualize (Khatwani and Srivastava, 2018). The TSNE objective function calculates how the "neighbors" of these similar data are distributed in two or three dimensions, and then maps them to the cluster. In previous studies, the objective minimization of TSNE was carried out as an n-body simulation problem, where points were randomly distributed in the embedding space and each point was affected by two different types of forces. Attraction pulls points closer and closer to their closest similarity in higher dimensional space, while repulsive forces push them further and further away from their embedded neighbors.

Customer Consumption Forecasts Based on Machine Learning Algorithms

Predictive analysis of customer spending and customer retention are extremely important research topics in CRM. In games, as well as other industries, predicting the amount of customers' consumption has great impact on the company's future profits. Among the main methods used are: logistic regression, decision tree, correlation analysis, neural network, SVM and random forest. With the development of technology, data mining now needs to process large amounts of data, and deal with missing data, abnormal data, strong randomness and many other issues. Each algorithm is not specifically designed to solve a particular problem, and algorithms are not independent or exclusive. For any problem, there is no such thing as the 'best' algorithm; various suitable models are debugged and run, and finally the best model is selected. The following sections give brief introductions to the common algorithms for solving customer consumption prediction problems.

Linear Regression

Linear Regression is an analysis that models the relationship between one or more independent variables and dependent variables using a least squares function. For a sample , the output value of the linear regression model is a linear combination of its characteristics. The goal is to fit the prediction result as much as possible to the dependent variable under actual conditions. The situation with only one independent variable is called simple linear regression, whereas where there are two or more independent variables it is called multiple linear regression.

Logistic Regression

Similarly, Logistic Regression is a machine learning method for solving the problem of two classifications (0 or 1), and estimate the possibility of a particular case. Both logistic regression and linear regression belong to a generalized linear model, but logistic regression assumes that the dependent variable, y, obeys the Bernoulli distribution, while the linear regression assumes that the

dependent variable y obeys a Gaussian distribution. At the same time, logistic regression introduces nonlinear factors through the Sigmoid function. The Sigmoid function is:

$$f(x)=\frac{1}{1+e^{-x}}$$

The value of the dependent variable of this function is , but its value range is (0,1). Therefore, regardless of the value of the independent variable, the value of the last Sigmoid function must fall between 0 and 1.

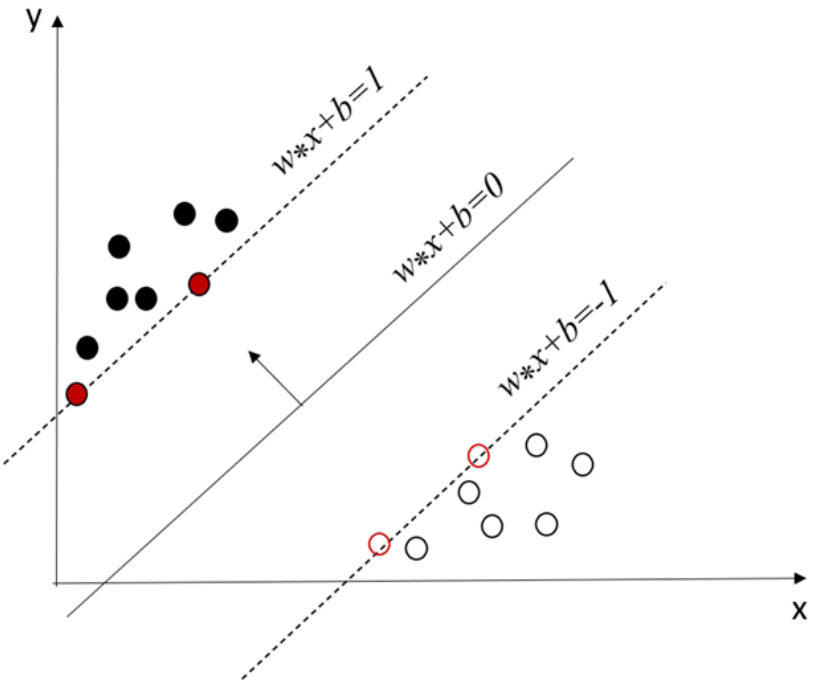
Decision Tree

The decision tree belongs to the supervised learning model category in machine learning. The posterior probability function *P(Y|X)* is used to obtain the probability of the category Y by the feature X. In general, the value of Y is a most classes (maximum likelihood estimation) record, after the eigenvector X is given. Decision tree includes feature selection, decision tree generation, and decision tree pruning. Common algorithms involve ID3, C4.5, CART regression trees, and CART classification trees. The most critical problem is the selection of optimal features. CART regression tree and CART classification also need to choose the optimal value of the feature.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is a two-class model, whose basic principle is to solve for a separate hyperplane that correctly divides the training data set and maximizes its geometric spacing. As shown in Figure 2, is the separation hyperplane. For linearly separable data sets, there are infinite

Figure 2. Schematic diagram of SVM hyperplane segmentation (Vapnik and Chervonenkis, 1964).

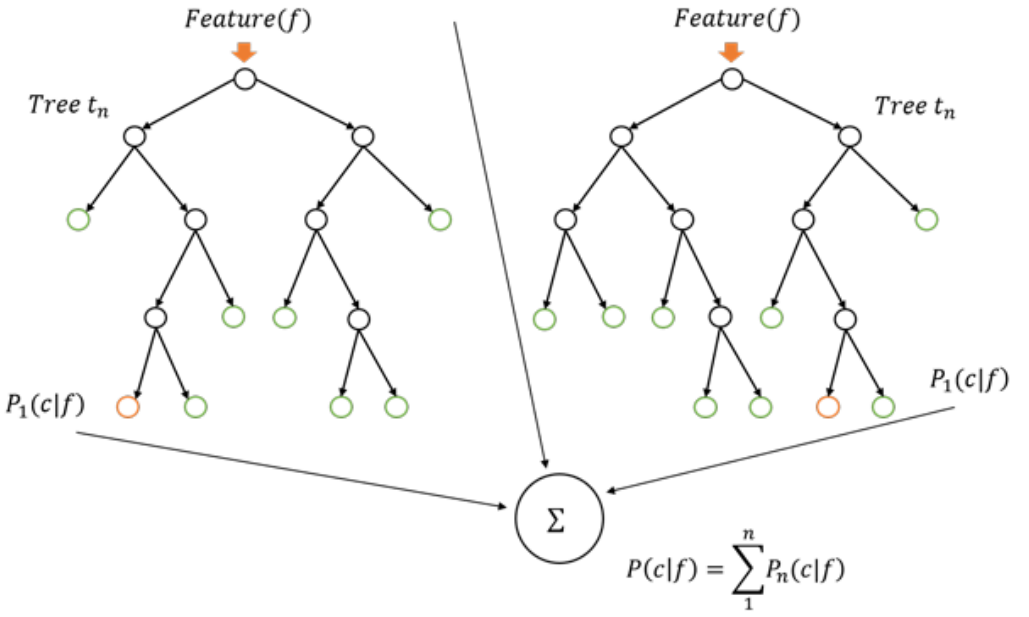


hyperplanes, but the separated hyperplane with the largest geometric spacing is unique. At the same time, SVM also includes nuclear functions, which enable it to become a virtually nonlinear classifier.

Random Forest

The concept of the 'Random Forest' was first proposed by Breiman (2001). It gives a set of tree classifiers , where the subclassifier is a fully grown classification regression tree established by the CART algorithm, and is the input vector, is an independent and identically distributed random vector. The random vector plays a decisive role in the growth of a single tree. The final output results in a simple majority voting method for classification problems and a simple averaging method for regression problems. The combination of Bagging algorithm and CART classification regression tree, together with the random selection of attribute splitting, gives the random forest a strong noise influence avoidance ability (Figure 3).

Figure 3. Principles of random forest classification (Zhao, Zhang and Geng, 2019).



RESULTS - CASE ANALYSIS OF DATA MINING TECHNOLOGY IN ONLINE GAMES

The structure of this case analysis is shown in Figure 4:

EMPIRICAL BACKGROUND

This study takes the whole player network of a game company in China as the research sample and predicts the total amount of consumption of game products and services in the future, based on its login and recharging data from June 23, 2017 to October 9, 2018. According to historical data analysis, the economic benefits of customers are mainly related to their stored value frequency, stored value time and stored value in the game, which is consistent with the three elements in the

traditional RFM model. Based on these points, this paper defines a new indicator S, which measures the volatility of customer's consumption behavior. The model is first used for predictive analysis of the amount of consumption for customers in game, and various indicators are measured by the entropy weight method. Each customer is scored by the value index; and then the K-Means algorithm is used to group the customers, in order to better specify the precise marketing solution for each type of customer. After obtaining the value index score of each customer, it was used together with other customer login and stored value information as an explanatory variable for predicting the customer's consumption. After comparing the prediction accuracy of multiple regression models, the random forest is selected as the optimal model. After optimization and adjustment, a customer consumption amount prediction model with higher prediction accuracy is obtained. Results of analyses are used to help the game company to rationalize the customer's personalized plan, determine the customer's stored value habits, and consolidate and develop customer relationships.

Model Construction

Sample Selection

To establish an RFMS model this study utilizes company customer login and stored value data from June 23, 2017 to October 9, 2018. This dataset has 8805 customers, each owns 34 explanatory variables. The variables are explained in Figure 5.

Related Indicator Analysis

According to the previously mentioned five indicators (DAU, MAU, SoU, ARPU, ARPPU) used to evaluate game management and player stickiness, Figure 6 illustrates the performance of the data.

Figure 6. Summary of performance of game data – Related Indicator Analyses:

The analysis of DAU shows that the most active time for daily game active users is from July to September of each year, and the number of daily active players in 2018 is higher than in 2017. The game had the most monthly active users in August 2017, but there was a significant reduction in users in August 2018. Analysis of SoU shows that user's adhesion is relatively low, with little change per month, although game adhesion in each month of 2018 is much higher than in 2017, indicating that the game's popularity among players has increased over time. Analysis of ARPU shows that the average income per user in 2017 is not as high as that of 2018, and the latter is very variable; it drops significantly in January, May, June, September and October. Analysis of ARPPU demonstrates that the average income per paying user is lowest in October 2017, and there was a large change in November 2017. During this data period, the average income per paying user was highest in July 2018.

RFMS Model

Data Standardization Processing

After pre-processing the basic data, the R, F, M, and S indicators of each customer are calculated. Since the dimensions of each indicator in the RFMS model are different, to eliminate the influence of large distribution differences and dimensions, the data are normalized before weighting each indicator. Among the four indicators, the R and S indicators are negatively correlated with the final customer value index, while M and F are positively correlated with the final customer value index. Accordingly, this study applies different standardized formulas for the two types of indicators.

For the F and M indicators, the following standardization process is used:

$$x' = \left(x - x^{s}\right) / \left(x^{s} - x^{t}\right)$$

where is the normalized value, is the original value, and is the minimum value, is the maximum value of the index.

For the R and S indicators, the normalization processing is:

$$x' = \left(x^l - x\right) / \left(x^s - x^l\right)$$

Figure 4. The structure of the case analysis

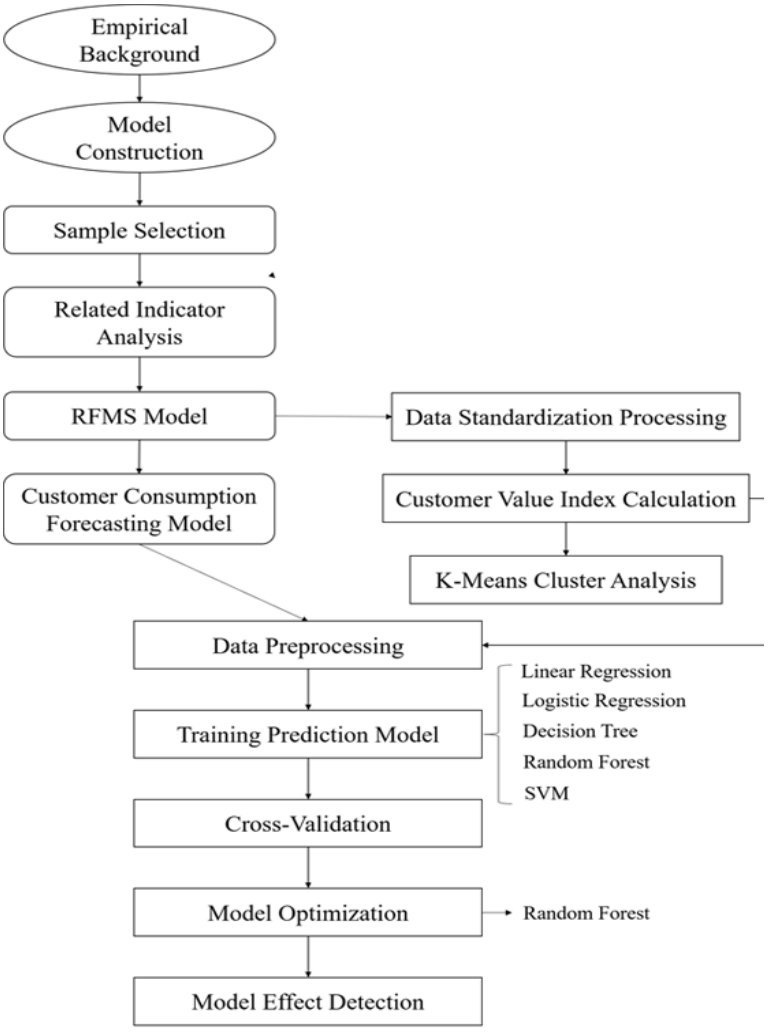


Figure 5. Summary of 34 Variable Explanations

Variable 1	Explanation 1	Variable 2	Explanation 2
AccountID	Customer ID	Pay5Sum	Total stored value of Pay 5
1LevelMean	Average level of 2nd transmigration	Pay6Mean	Average stored value of Pay 6
2LevelMean	Average level of 2nd transmigration	Pay6Sum	Total stored value of Pay 6
CarrerIDCount	Number of occupations	Pay7Mean	Average stored value of Pay 7
MoneySum	Customer's current amount	Pay7Sum	Total stored value of Pay 7
NewlevelMean	Average level of transmigration	Pay8Mean	Average stored value of Pay 8
NewtimesSum	Times of transmigration	Pay8Sum	Total stored value of Pay 8
PartMean	Average number of partners	RepuMean	Average reputation
Pay1Mean	Average stored value of Pay 1	RepuSum	Total reputation
Pay1Sum	Total stored value of Pay 1	SexMean	Average gender (0:male; 1:female)
Pay2Mean	Average stored value of Pay 2	MeanloginDay	Frequency of login
Pay2Sum	Total stored value of Pay 2	SumLoginTimes	Total logins
Pay3Mean	Average stored value of Pay 3	CountTime	Total stored times
Pay3Sum	Total stored value of Pay 3	MeanTime	Frequency of storing
Pay4Mean	Average stored value of Pay 4	StdMoney	Standard division of store value
Pay4Sum	Total stored value of Pay 4	SumMoney	Total stored value
Pay5Mean	Average stored value of Pay 5	NearestPayTime	Nearest storing date

Figure 6(a). DAU

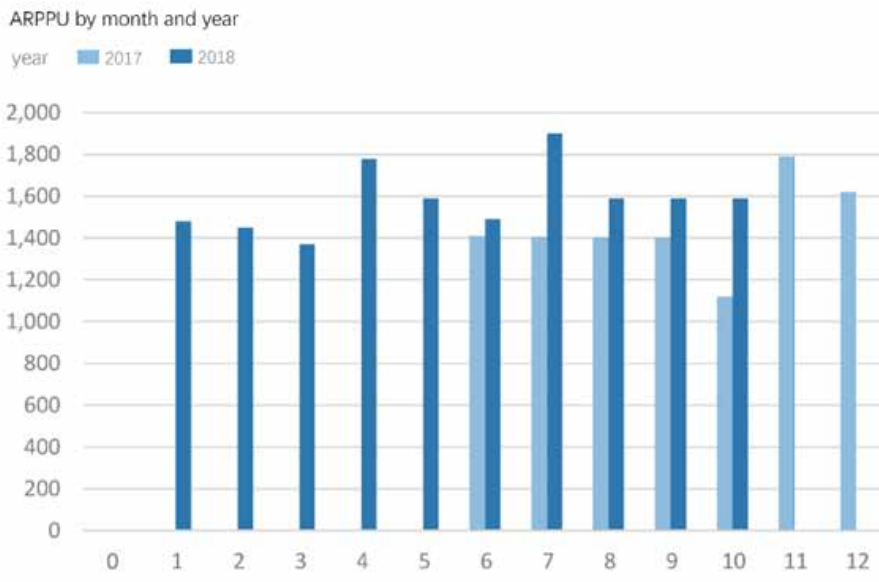


Figure 6(b). MAU

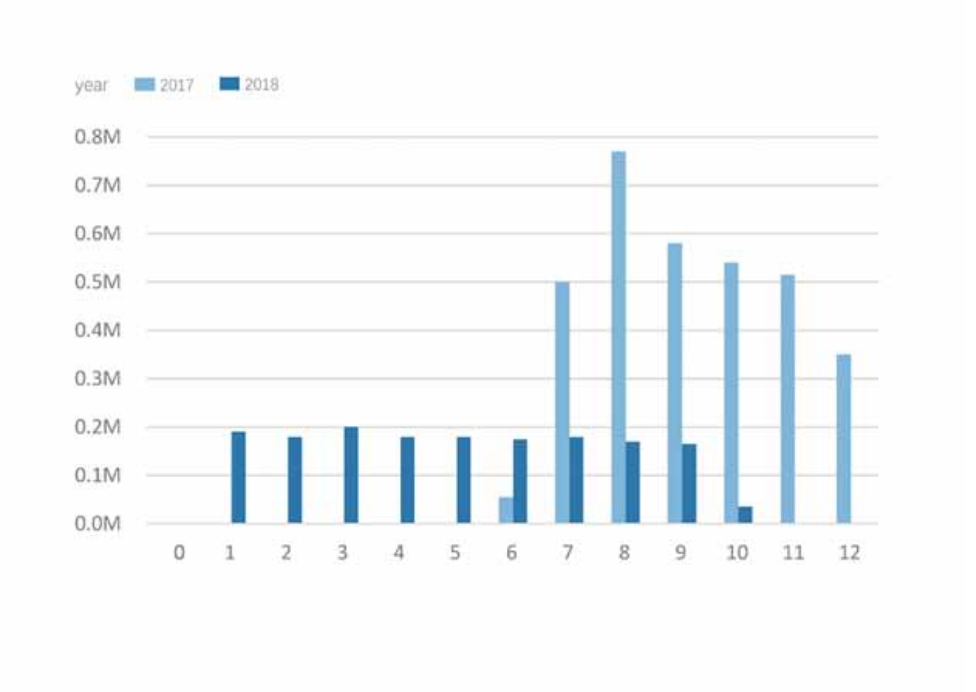


Figure 6(c). SoU

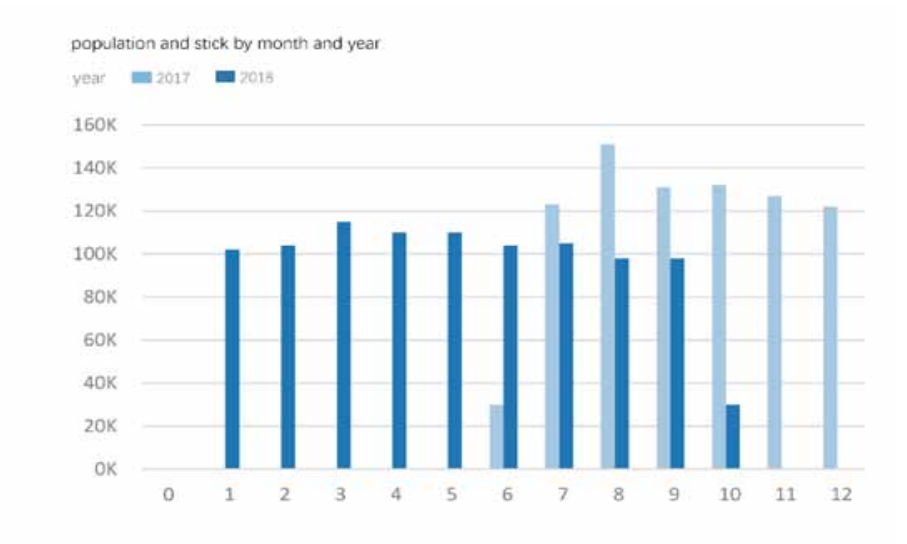


Figure 6(d). ARPU



Figure 6(e). ARPPU.

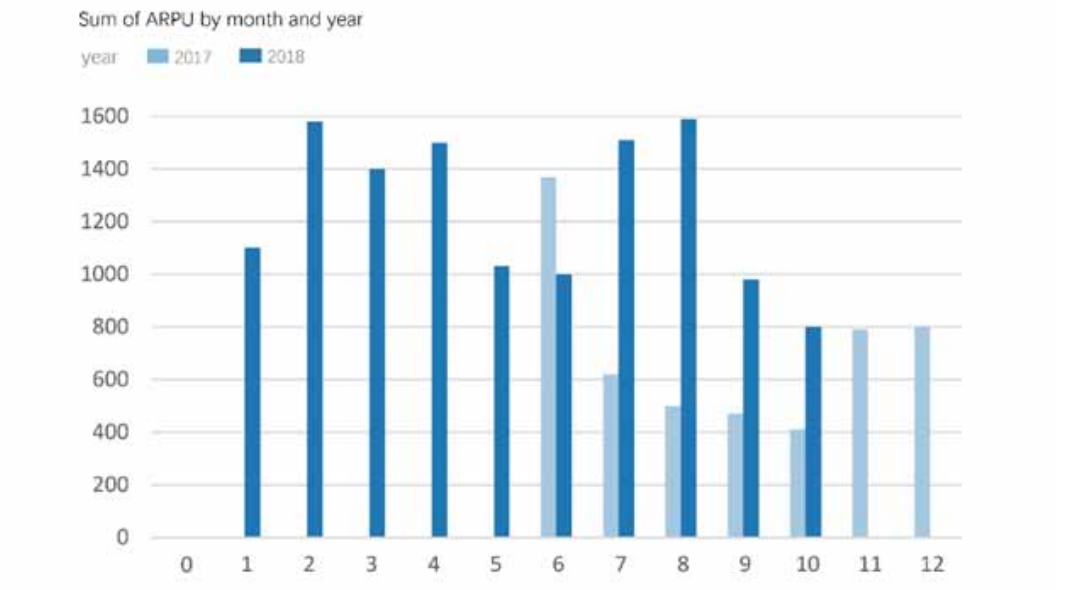
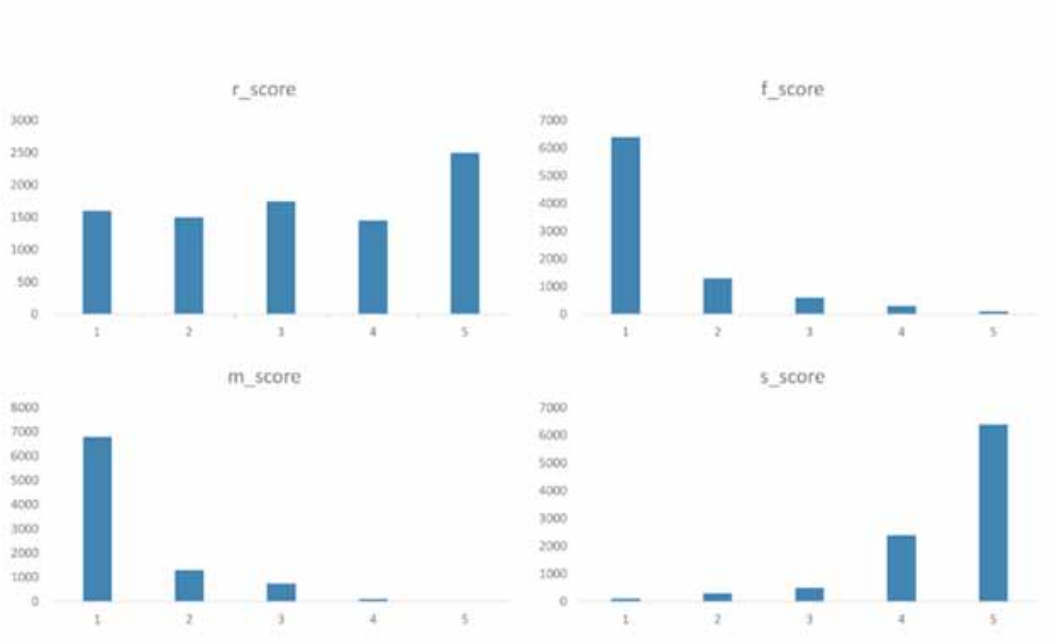


Figure 7 shows that among the four indicators, most of the customers' F, M, and S indicators are concentrated, distributed at 1 point, 1 point, and 5 points respectively. This indicates that most customers have low stored value frequencies and small total stored value, with good standard deviation of stored value, indicating that the customer's consumption behavior is less volatile; at the same time, the customer's performance on the R indicator is not the same. Between 1 and 4 partitions, the number of customers is similar, and the number of customers with 5 points is significantly higher. This indicates that most of the customers are still new players, and the company can continue to tap the consumption potential in the future.

Figure 7. Distribution of indicators in RFMS models for all customers



Customer Value Index Calculation

This index is a comprehensive index of the RFMS model, with a distribution range of 1 to 5 points. The higher customer's value index, the greater the customer's economic value to the company, and vice versa. After using the RFMS model to obtain the indicators of clients, the entropy weight method is then used to determine the weight of the four model indicators, in order to measure the impact of each indicator on the customer value index more accurately. The higher the weight, the greater the contribution of the indicator to the customer's value index. The game company can then consider opening up more markets for high value index customers. The weights of the indicators obtained by the entropy weight method are summarized in Figure 8.

Figure 8. Weights of indicators in the RFMS model

Indicators	Weight
R	0.02924701
M	0.44586952
F	0.52469507
S	0.00018839

The formula for calculating the customer value index is as follows:

$$\text{Score} = \beta_1 \cdot \text{R} + \beta_2 \cdot F + \beta_3 \cdot M + \beta_4 \cdot S$$

After getting the value index for each customer, the distribution of value indices across all sample data is shown in Figure 9.

Figure 9 shows that the value index distribution of each customer is relatively uneven. The value for most customers is concentrated in the 1.0 -1.5 interval, with only a few customers distributed in the 4.5 - 5.0 range. This means that for this game, there are fewer customers with higher economic value, who are actually the group that brings the most profit to the company.

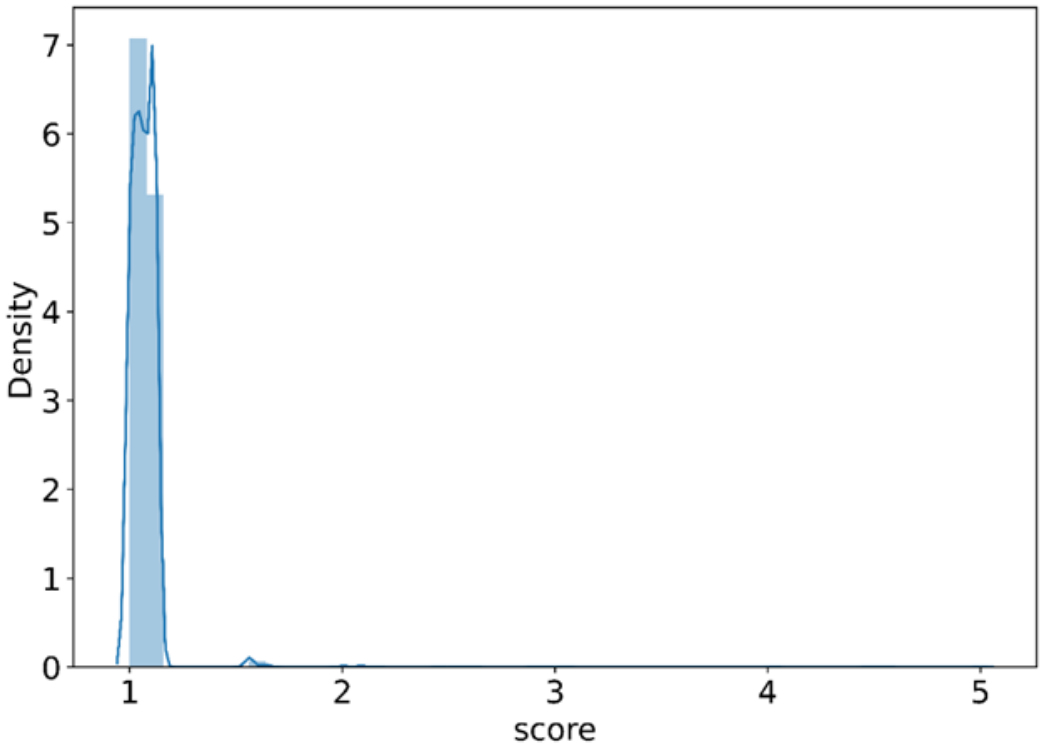
K-Means Cluster Analysis

Following the standardized scores of four indicators and the value index, this research continues to use the K-Means algorithm to group customers and the TSNE algorithm to reduce the data dimension.

Before using K-Means for clustering, it is necessary to pre-determine the total number of categories of clusters. In this model, the customer classification depends on the comparison of four indicators R, F, M, and S of each customer, with the average value of all samples. For example, if the R indicator of customer A is higher than the average of the R index of all customers, the R indicator of customer A is defined as , and if the R indicator of customer A is lower than the average R value of all customers, the R indicator of customer A is defined as ¯. In this way, the four indicators are determined.

Accordingly, when applying this method to define each indicator category of the customer, we first divided all customers into 2×2×2×2=16 categories. However, following consideration of the convenience and visibility of the company's customer classification in practice, as well as the number of customers in each category, the initial 16 customer groups were integrated into 8 categories for better comparison and analysis - based on the number of up and down arrows of the four indicators in each customer category. These eight categories were: important value (IV), important maintenance

Figure 9. Distribution of all customer value indices



(IM), important development (ID), important retention (IR), general value (GV), general maintenance (GM), general development (GD) and general retention (GR). The levels of each customer cluster are shown in Figure 10.

Customer classification not only reveals the difference in the level of each type of customer, but also reflects the behavioral characteristics and changing tendencies of the customer. By classifying existing customers, game companies can adopt different management and marketing strategies for different levels of customers.

Customer Consumption Forecasting

Based on the basic data and the calculated customer value index, this study identifies 34 explanatory variables and one explained variable (total stored value). With these, it is hoped to construct a model that predicts the amount of future customer spending.

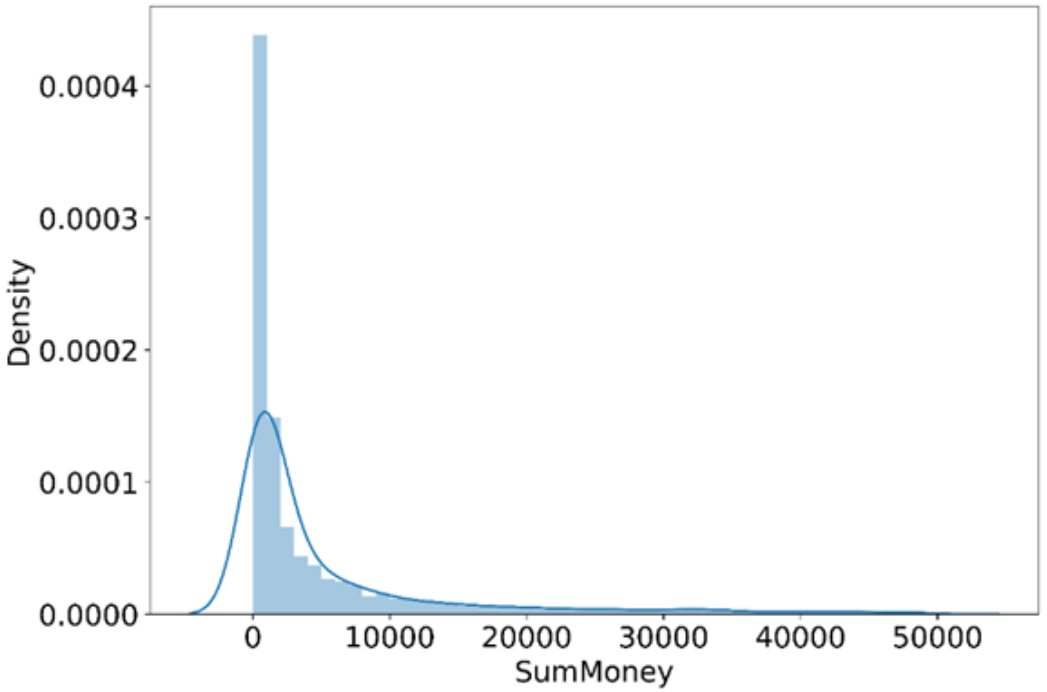
Figure 10. Customer category segmentation

Num	Rate	Zr_score	Zf_score	Zm_score	Zs_score	Score	Feature	Category	CateNum
1	0.06%	1.2278	23.1983	33.1142	-2.6198	27.7537	↑ ↑ ↑ ↑	IV	5
2	0.20%	1.1908	6.7676	10.6975	-0.64	8.6651	↑ ↑ ↑ ↓	IM	18
3	0.62%	1.2036	8.1368	-0.0502	0.1214	3.6368	↑ ↑ ↓ ↑	ID	55
4	0.25%	0.501	-0.0785	3.7185	-14.2078	1.928	↑ ↓ ↑ ↓	IR	22
5	27.50%	1.2278	-0.0785	-0.0502	0.1214	-0.0254	↑ ↓ ↓ ↑	GV	2421
6	1.57%	0.6388	-0.0785	-0.0502	-5.1426	-0.0437	↑ ↓ ↓ ↓	GM	138
7	32.58%	0.1787	-0.0785	-0.0502	0.1214	-0.0561	↓ ↓ ↓ ↑	GD	2869
8	37.22%	-1.1224	-0.0785	-0.0502	0.1214	-0.0942	↓ ↓ ↓ ↓	GR	3277

Data Preprocessing

First, a basic data description of 34 known explanatory variables, and the distribution state of each variable is provided. Due to the small number of customers (8805), a large error could be produced if traditional random sampling is used to form training samples and test samples. Therefore, stratified sampling was adopted for the segmentation of samples, which is most appropriate where there is sample imbalance. Stratified sampling divides the population into disjoint levels, and extracts a certain number of individuals independently from each layer according to a certain ratio. The individuals taken out of each layer are combined as a training sample and test sample. Figure 11 shows the distribution of the interpreted variables for all samples in the data source.

Figure 11. Distribution of total customer spending



Having formed the training set and test set, the correlation between each explanatory variable and the interpreted variable is calculated. The first six variables were selected according to the size of the correlation coefficient, and the Pearson correlation coefficient graphs plotted (Figure 12).

In addition to the six explanatory variables mentioned above, according to empirical practice, all explanatory variables with correlations to explained variables of less than 0.2 were removed, leaving only 14 explanatory variables to train the model and improve prediction ability. After screening the independent variables, the characteristics of these variables were further scaled, using the same standard to measure the predictive ability of all variables and reduce prediction error.

Training Prediction Model

Based on the above calculations, linear regression, logistic regression, decision tree, SVM and other models can be applied to predict the customer's consumption amount. The investigation trained

multiple models for the training set data and measured the prediction error of each model with Mean Square Error (MSE). The smaller the MSE, the better the performance of the predictive model in describing the experimental data, the stronger the predictive ability, and vice versa. The mean square error of the above five models on the training samples are shown in Figure 13.

Figure 12. Pearson coefficient plot of six important variables

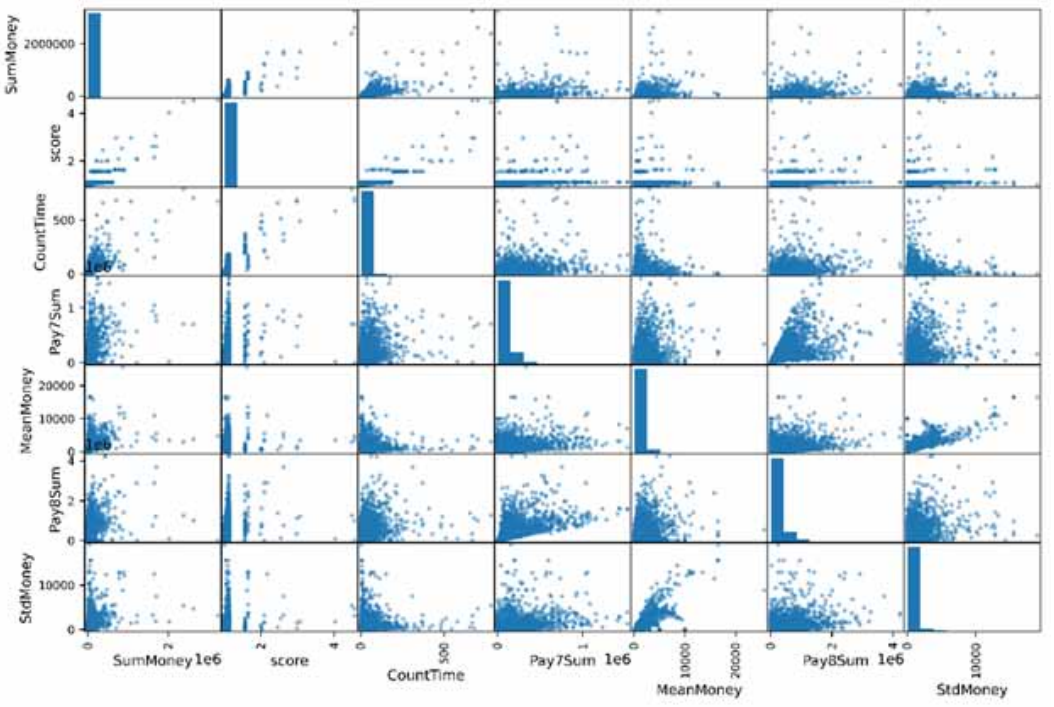


Figure 13. Comparison of MSE for each model

Model	MSE
Linear Regression	41698.3677
Logistic Regression	38275.6697
Decision Tree	11552.2386
Random Forest	10387.3101
SVM	3905.3970

Comparing the mean square error of each model on the training set, it can be seen that the MSE of linear regression and logistic regression is large. This indicates that the performance of these two models is poor, while the performance of decision trees and random forests is moderate. It is clear that SVM has the smallest MSE, indicating that this model is the best at present; however, because there is only one training result the robustness of the model cannot be determined, and an over-fitting situation cannot be ruled out. Accordingly, 10 cross-validations on 5 models respectively were carried out, in an attempt to obtain the model with the smallest prediction error through multiple estimations.

Cross-Validation

Cross-Validation, is a practical method of cutting data samples into smaller subsets. The model analyzes a subset first, while the other subsets are used as confirmation and verification entities. The ultimate goal is to reduce the occurrence of over-fitting and instability of the model.

The above five models were cross-validated, and the final MSE mean of each model obtained; this value refers to the average MSE after 10 calculations of the original model, and the standard deviation of each model in 10 trainings is also obtained. It can be considered as an indicator to measure the volatility of the model. Figure 14 shows the MSE results for cross-validation of each model.

As the figure shows, the random forest model performs best, with the lowest average MSE. While the average MSE of the best performing SVM model in the first prediction is higher than that of random forest, indicating that in the first prediction, the SVM model does have an over-fitting problem; so, the MSE for the first training is only 3903.397. Based on the characteristics of lower average MSE and lower forecasting volatility, we finally chose the random forest model as our consumption amount prediction model.

Figure 14. Average MSE comparison of each model after cross-validation

Model	Average MSE	Standard Deviation
Linear Regression	29570.3605	5464.3782
Logistic Regression	11697.6607	24046.5823
Decision Tree	13990.6238	5901.2527
Random Forest	8995.7014	5845.6339
SVM	11924.5539	25002.1412

Model Optimization

After determining the model, due to the fact that parameters in the random forest model are manually adjusted with error to some extent, we used the GridSearch method to optimize the model; this ensured that the model's predictive ability reached its peak, and reduced error as much as possible.

GridSearch is a comprehensive method for specific parameter values. By optimizing the parameters of the estimation function through cross-validation, the optimal learning algorithm is obtained; this means the possible values of each parameter are arranged, combined and listed. All

possible combined results generate a "grid", each used for model training and evaluated based on cross-validation. After the fitting function has tried all the parameter combinations, it returns a suitable classifier and automatically adjusts to the optimal parameter combination. Through GridSearch, the optimal parameters of the random forest model were easily determined and the characteristic important values of the 14 independent variables calculated. The eigenvalues of variables permit us to improve the performance of the model, and also allow identification of the characteristic variables that have the greatest impact on the player's total stored value in the marketing plan; they also improve the company's understanding of the importance of these characteristic variables.

Model Effect Detection

After training the model and adjusting parameters, the efficiency of the model was tested. First, the test set data were processed, variables with less influence on the total value of the player's stored value (correlation less than 0.2) were removed; then feature scaling and normalization conversion on the remaining variables was performed, to obtain the standard test set format before putting into the model. Then, using our final optimization model, we predict the data of the test set, and output the prediction results and errors. The final MSE is 15781.8259. This error is greater than from the training set, but this situation is normal (the error of the general test set is usually greater than that from the training set); it also indicates that this optimization model needs further refinement.

DISCUSSION AND CONCLUSION

This paper highlights the significance of the customer value index and consumption forecasting for the game industry; it summarizes previous research results and combines them with the characteristics of the industry. To the traditional RFM model, an S indicator is added, with the aim of constructing an RFMS model that more fully measures the customer's value index. The regression models used in the forecasting of customer consumption can then be re-interpreted.

This study utilizes player data provided by a game company in China over a 15 month period. Based on login and stored value data, we use five professional indicators commonly used in the game industry to count and analyze the data, and obtain the distribution structure of all customers. This also yields a relatively comprehensive understanding of the game operation and customer login information.

To develop a more comprehensive model this study follows a process of general data mining; successive phases in the process include: data processing, feature engineering, model building, effect evaluation and parameter optimization. The improved RFM model was applied first to the customer on the calculation of each index, using the entropy weight method to determine the weight of the four indicators, so building a value index that could evaluate the customer's economic benefits to the company. Then all customers were divided into 8 groups, according to the K-Means clustering method, depending on the scores of the four indicators. This segmentation shows that customers in category 1 and category 2 have high economic value for the game company, but their total number is only 23, which is a very low percentage of the total customers. In addition, categories 3 and 4 closely follow, where value index is relatively lower than that of the first two types of players. Groups 3 and 4 consist of important development customers and important retention customers, respectively; then category 5 and category 6 account for the largest proportion of customers, which is the main development target of the game company in the future. If the marketing strategy turns these customers into important value customers, then the company will create excess profits. The final category 7 and 8 customers have lower value, and the company should focus on improving the adhesion of these customers to the game.

Next, linear regression, logistic regression, decision tree, SVM and other models are applied to predict the future consumption amount of customers respectively. After comparing the effectiveness of each model by means of cross-validation, we finally chose random forest as the customer consumption

prediction model. The GridSearch method was used to optimize the model for predicting the customer's consumption amount.

In practical terms, this research has significance for the processing of player data of the game company and the establishment of predictive models. It illustrates economic benefits for game companies adopting this process, including but not limited to: marketing strategies for different categories of players; more accurate formulation and understanding of new players' stored value habits; more accurate prediction of new players' future spending; and more precise targeting of new packages or offers, that provide incentives for new players to recharge and increase the likelihood of multiple purchases in the future. In addition, this development study applies and improves the model in a game-specific industry, based on fundamental models; however, the overall analysis process can also be applied in other industries, and play a role in corporate marketing elsewhere.

After considering the data and methods used, it is apparent that this study still has limitations. First, due to the perceived value and sensitivity of data, we were not able to obtain multiple customer consumption data from multiple game companies. The analysis can only be based on data from a single company game, thus there are not enough training samples. The error of the customer consumption prediction model obtained at the end of this paper cannot be further reduced.

Second, the customer's original data provided is complicated, but the information covered is limited. For example, there was a lack of information on specific products and services of the player in the game, as well as the relation between game sets, and basic information on game products. This resulted in a situation where we couldn't fully analyze all the game data of customers, nor could we more accurately evaluate the subjective factors, such as the customer's preference for the game product, or the psychological consumption expectations.

With the advent of artificial intelligence, all types of industry are facing a major revolution. In the game industry, companies will incorporate big data technology to strengthen customer management, improve customer portraits, enrich customer labels through data mining, conduct customer classification, and achieve internal and external data interconnection. With the higher quality of games and users' growing attention, the traditional model of 'simply producing a good game' is no longer applicable. At this stage, the market strategy appears to be the essential key for success. From the perspective of consumers, the taste and screening ability of a game is getting stronger and stronger. Only more accurate recommendations and valued content can meet the needs of players. From the manufacturer's point of view, with the strong rise of 'hard core alliance', the industry chain of game distribution is gradually self-upgrading; but this means that the value of game lists is getting lower and the value of player word-of-mouth and marketing is becoming more important. These are inevitable business and development trends.

The significance of this study is not only to add a dimension, but also contribute to poorly known aspects of the game field. The overall objective is, relying on the big data of the game industry, to inspire the company to think about the customer business model and apply big data analysis to all kinds of business services. In future, it will become a new entry point for the transformation to more sustainable business management.

REFERENCES

Beckschafer, P., Fehrmann, L., Harrison, R. D., Xu, J., & Kleinn, C. (2014). Mapping Leaf Area Index in subtropical upland ecosystems using RapidEye imagery and the random Forest algorithm. *Iforest - Biogeosciences and Forestry*, *7*(1), 1-11.

Behrens, C., Pierdzioch, C., & Risse, M. (2018). Testing the optimality of inflation forecasts under flexible loss with random forests. *Economic Modelling*, *72*, 270–277. doi:10.1016/j.econmod.2018.02.004

Benoit & Van den Poel. (2010). Random forests. *Machine learning*, *45*(1), 5-32.

Chen, Y., Cheng, C., Chen, D., & Chen, W. (2012). Applying Feature Selection Combination-Based Rough Set Classifiers to Forecast Credit Rating Status. *International Conference on Genetic and Evolutionary Computing*, 425-428. doi:10.1109/ICGEC.2012.67

Cheng, C.-H., & Chen, Y.-S. (2008). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3), 4176–4184. doi:10.1016/j.eswa.2008.04.003

Chia-Huei Wu (PhD) is the corresponding author of this project. They are currently an Assistant Professor at Institute of Service Industries and Management, Minghsin University of Science Technology in Taiwan.

Ciner, C. (2019). Do industry returns predict the stock market? A reprise using the random forest. *The Quarterly Review of Economics and Finance*, *72*, 152–158. doi:10.1016/j.qref.2018.11.001

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives*, *18*, 153–160. doi:10.1016/j.tmp.2016.03.001

Fabisiak, L. (2018). Web service usability analysis based on user preferences. *Journal of Organizational and End User Computing*, *30*(4), 1–13. doi:10.4018/JOEUC.2018100101

Feng, Y., & Wang, S. (2017). A forecast for bicycle rental demand based on random forests and multiple linear regression. *Annual ACIS International Conference on Computer and Information Science*, 101-105. doi:10.1109/ICIS.2017.7959977

Gao, J., & Lu, X. (2015). Forecast of China railway freight volume by random forest regression model. *International Conference on Logistics Informatics and Service Sciences*, 1-6.

Gartner, R. (2014). A metadata infrastructure for the analysis of parliamentary proceedings. *International Conference on Big Data*, 47-50. doi:10.1109/BigData.2014.7004452

Gounaridis, D., & Koukoulas, S. (2016). Urban land cover thematic disaggregation, employing datasets from multiple sources and RandomForests modeling. *International Journal of Applied Earth Observation and Geoinformation*, *51*, 1–10. doi:10.1016/j.jag.2016.04.002

Herrera, G. P., Constantino, M., Tabak, B. M., Pistori, H., Su, J.-J., & Naranpanawa, A. (2019). Long-term forecast of energy commodities price using machine learning. *Energy*, *179*, 214–221. doi:10.1016/j.energy.2019.04.077

Hobley, E. U., Baldock, J., & Wilson, B. (2016). Environmental and human influences on organic carbon fractions down the soil profile. *Agriculture, Ecosystems & Environment*, *223*, 152–166. doi:10.1016/j.agee.2016.03.004

Hsieh, N.-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, *27*(4), 623–633. doi:10.1016/j.eswa.2004.06.007

Kasiran, Z., Ibrahim, Z., & Ribuan, M. S. M. (2014). Customer churn prediction using recurrent neural network with reinforcement learning algorithm in mobile phone users. *International Journal of Intelligent Information Processing*, *5*(1), 1.

Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. *Procedia Computer Science*, *3*, 1327–1332. doi:10.1016/j.procs.2011.01.011

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57–63. doi:10.1016/j.procs.2010.12.011

Khatwani, G., & Srivastava, P. R. (2018). Impact of information technology on information search channel selection for consumers. *Journal of Organizational and End User Computing*, *30*(3), 63–80.

Khobzi, H., Akhondzadeh-Noughabi, E., & Minaei-Bidgoli, B. (2014). A new application of rfm clustering for guild segmentation to mine the pattern of using banks' e-payment services. *Journal of Global Marketing*, *27*(3), 178–190.

Lessmann, S., & Voß, S. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting*, *33*(4), 864–877. doi:10.1016/j.ijforecast.2017.04.003

Lohrmann, C., & Luukka, P. (2019). Classification of intraday S&P500 returns with a Random Forest. *International Journal of Forecasting*, *35*(1), 390–407. doi:10.1016/j.ijforecast.2018.08.004

Ravasan, A. Z., & Mansouri, T. (2015). A fuzzy ANP based weighted RFM Model for customer segmentation in Auto Insurance sector. *International Journal of Information Systems in the Service Sector*, *7*(2), 71–86. doi:10.4018/ijisss.2015040105

Shahri, A., Hosseini, M., Phalp, K., Taylor, J., & Ali, R. (2019). How to engineer gamification: The consensus, the best practice and the grey areas. *Journal of Organizational and End User Computing*, *31*(1), 39–60. doi:10.4018/JOEUC.2019010103

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Tamaddoni Jahromi, A. (2009). *Predicting customer churn in telecommunications service providers* (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-46732

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, *79*, 679–683. doi:10.1016/j.procir.2019.02.042

Toffler, A. (1980). *The Third Wave*. Bantam Books.

Vapnik, V., & Chervonenkis, A. (1964). A note on class of perceptron. *Automation and Remote Control*, 24.

Woodruff, R. B. (1997). Customer value: The next source for competitive advantage. *Journal of the Academy of Marketing Science*, *25*(2), 139–153. doi:10.1007/BF02894350

Zhao, Y., Zhang, Y., Geng, M., Jiang, J., & Zou, X. (2019). Involvement of slab-derived fluid in the generation of Cenozoic basalts in Northeast China inferred from Machine Learning. *Geophysical Research Letters*, *46*(10), 46. doi:10.1029/2019GL082322

Zilin Deng is a senior majoring in finance (securities and futures) in Southwestern University of Finance and Economics. Their research direction is quantitative trading.

Zongxiao Wu is a student of Finance School at the Southwestern University of Finance and Economics. Wu's interests are focused on the measurement and management of credit risk including credit scoring, credit rating, and risk modelling under Basel II and III, though her research mainly employs big data in financial risk management. Cong Zang is a researcher currently in Southeast University, Nanjing, China, who majors in Electronic Science and Engineering.

Xuefeng (David) Shao (PhD) is a lecturer in management at The University of Newcastle. He earned his PhD degree in International Business from UNSW and was a research fellow at UTS. His research has appeared in different A journals on ABDC list including Technological Forecasting & Social Change, Journal of Cleaner Production, Journal of Environmental Management, Enterprise Information Systems, Journal of Global Information Management, Pacific-Basin Finance Journal, and Finance Research Letters.

Wei Liu is a Distinguished Professor at Business School, Qingdao University. His current research interests include international business and corporate strategy in the Chinese context. He has published several papers in refereed journals and presented numerous papers at top-tier international business conferences. He also served as the reviewer in many journals and conferences. He has won two finalists of the AIB and AOM best paper.