Financial Engineering and Risk Management (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2023.060816
ISSN 2523-2576 Vol. 6 Num. 8

Research on the Risk Analysis of Cigarette Business Based on Data Mining Technology

Senqiang Wang*, Xiao Song

School of Management, Shandong University of Technology, Zibo, 255000, China
**Corresponding author*

Keywords: Data mining, cigarette management, risk analysis, sk-learn algorithm

Abstract: It has always been the focus of the tobacco industry to carry out the risk investigation of standardized tobacco management. At present, there are still large risk management problems in the links of brand supply, supply distribution of retail households, real cigarettes and acquisition cases, which destroys the coordination of cigarette sales market. With the help of big data, we can effectively identify and supervise illegal business risks, so as to reduce the occurrence of abnormal situations. Based on the data mining technology, this paper comprehensively uses the supervised sk-learn algorithm, extracts and transforms the data analysis index in the cloud pos system, and studies the abnormal data in the process of cigarette business. In addition, this constructs a risk analysis model for cigarette business, which promotes the data governance and risk control ability of the cigarette market, and is conducive to optimizing the overall image of the tobacco industry.

1. Research Background

Over the years, the State Bureau, in accordance with the Tobacco Monopoly Law, has established a tobacco monopoly system, aiming to regulate the production and operation of cigarettes in the tobacco industry. Under this system, although the operation of tobacco commercial enterprises has maintained a stable level of control, the violation of laws and regulations is repeatedly banned due to the lack of data and systematic management mode. In addition, influenced by the internal and external environment of the industry, the tobacco industry is also facing different types and different degrees of operational risks.

There are many reasons for this phenomenon, mainly including the following two points: On the one hand, in order to regulate the production and operation activities of the tobacco industry, the Tobacco Monopoly Administration has always maintained a highly strict working situation, and various localities have issued various regulations and documents on smoking ban. This makes the network terminal household cigarette sales performance appeared sharply decline, leading to the increase of cigarette inventory and other problems. In addition, some retailers sell the stock of cigarettes intensively in order to seek economic benefits, thus disrupting the order of the cigarette market and greatly improving the risk of cigarette operation. On the other hand, due to the differences in the consumption degree and market demand of various cigarette brands in different regions, some businesses often purchase a large number of cigarettes hoarded by retail households in the warehouse and resell them to other consumption areas, which increases the difficulty of

standardizing the business operation of the tobacco industry.

With the deepening of monopoly management, the requirements of national and provincial bureaus for market supervision and standardized operation of enterprises have been gradually increased. Despite certain research results and implementation plans, there is still no comprehensive collection of abnormal internal operation data from the root, so as to achieve real-time risk early warning work. Therefore, with the help of data mining technology and hierarchical analysis method, this paper specifically constructs the operation risk identification and supervision system of retail households to curb the occurrence of business risks from the source.

2. Research Ideas and Goals

2.1. Clarify the Data Analysis and Transformation Process

Because the illegal operation of cigarettes involves many related parties, the reasons for the formation are even more complex and extensive, so it is impossible to comprehensively collect the key features and relevant information only by relying on the traditional working methods and experience. Therefore, the more ideal data analysis process should be to comprehensively collect all kinds of data[1], from which the data is uniformly classified, and extract the general rules and patterns of the data, and then adopt effective algorithms and models to process and judge the data set, and give field verification and risk supervision. Therefore, the whole data analysis and transformation process should be divided into two parts: "data collection and processing" and "model building".

2.2. Data Mining Based on the Cloud Pos System

Based on the cloud pos system, mining the relevant data of consumers consuming cigarette products. In order to ensure the comparability of data, taking into account the strict constraints of cigarette marketing business and marketing business needs, according to the difficulty of data processing, cloud pos data collection and processing are divided into days, weeks and months, which are determined according to the actual requirements. At the same time, attention should be paid to the seasonal and holiday volatility of cigarette sales. Consumer data within a cycle should be taken as timing data, including maximum, minimum, average and other data indicators, as important data samples for follow-up index analysis.

2.3. Analysis and Measurement of Multidimensional Key Features

On the one hand, tobacco monopoly, marketing, internal management and other departments have collected many business type indicators such as orders, brands, logistics and cases in their daily work. By classifying and analyzing the index information one by one, and systematically summarizing the analysis results of each indicator, the indicators that can reflect the cigarette business status are screened out, so as to determine the hierarchical evaluation criteria, and then initially identify the various risks in the process of production and operation. On the other hand, according to the calculation and evaluation results of each index system, the risk is further measured and estimated[2]. On the basis of measuring the degree of risks, the key characteristics and control strategies of risks are defined, and a grid operation supervision system is constructed, thus guiding the establishment of risk early warning system.

3. Study Design

3.1. Algorithm Selection

The accurate construction and utilization of the model is inseparable from the support of the algorithm. Different from the traditional programming technology, machine learning algorithm, as an efficient artificial intelligence technology, can effectively solve the problems of data classification, regression and clustering, and strengthen the data analysis and modeling capability when applying data mining technology[3]. Therefore, in this paper, based on the sk-learn moduleprovided in the jupyter notebook development environment[4], we conducted supervised machine learning (i. e., establishing labels for each sample in the target), and established a model using decision tree, Naive Bayes algorithm, logistic regression and k-nearest neighbor. Further, by integrating the predictive values of multiple models, more accurate prediction effects are obtained and evaluated using four indicators: accuracy, precision, recall and f1 score. Its application context can be orderly delimited into six major processes (as shown in Figure 1).

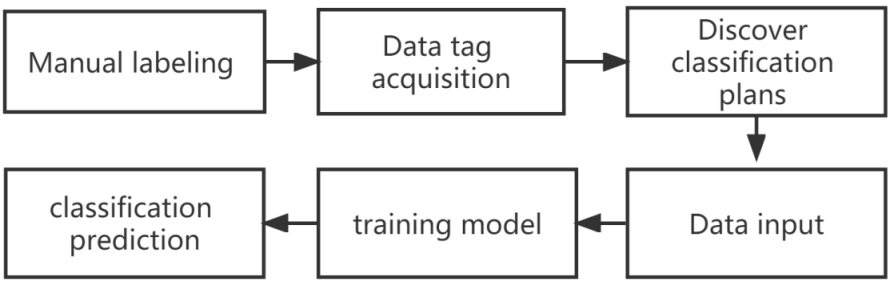

Figure 1: Application context of machine learning

3.1.1. Decision Tree

Decision tree algorithm is the most common supervised classification algorithm in data mining technology. Its principle is to select an attribute from all the existing conditions as the root node, and then look down whether there are other judgment conditions, and divide them into internal nodes or leaf nodes[5]. In the final decision tree, all leaf nodes are the category information to be output, while the root node and internal nodes are both feature information (as shown in Figure 2).

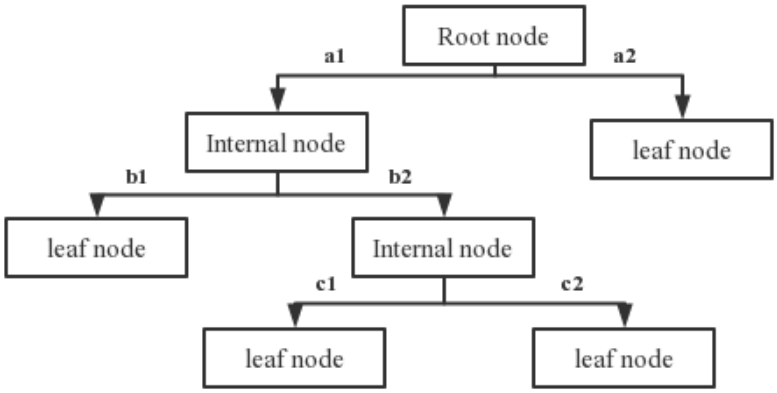

Figure 2: Structural diagram of the decision tree algorithm

In addition, in order to ensure the optimal node and feature classification ability of the decision tree, we also introduce the method of information gain and Gini index to measure the importance between features[6], calculating the impurity, and solve the problem of continuous value classification.

3.1.2. Naive Bayes Algorithm

Naive Bayes algorithm is also one of the important algorithms of data mining technology, which is a method that assumes the independence of different feature conditions[7]. It first takes the feature words in the given training set as the premise assumption, and then learns the joint probability distribution from input to output, and then the input b finds the output a that maximizes the posterior probability.

The expression formula is as follows:

$$P(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P}(\mathrm{B}|\mathrm{A})\mathrm{P}(\mathrm{A})}{\mathrm{P}(\mathrm{B})} \tag{1}$$

Where a is the categorical category and b is each feature attribute.

3.1.3. Logistic Regression

Logistic regression Is a commonly used dichotomization model[8], which can be used to predict whether the customer will buy a certain cigarette, or the possibility of buying a certain cigarette, so as to find the risk factors affecting business activities, predict and judge its occurrence probability[9]. The essence is to assume that the data obey this distribution and then use maximum likelihood estimates.

Its functional form is as follows:

$$g(z) = \frac{1}{1+e^{-z}} \tag{2}$$

Where e is the natural logarithm, with an infinite non-cyclic decimal.

The logistic regression formula diagram is shown in Figure 3:

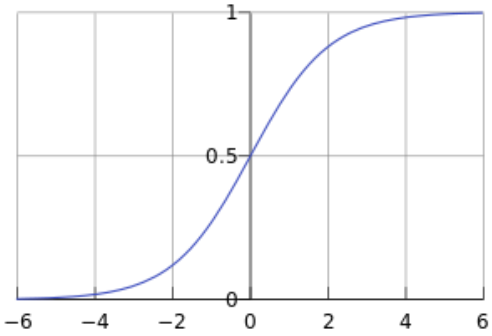

Figure 3: Loglogistic regression formula

3.1.4. K-Nearest Neighbor

K-nearest neighbor is one of the supervised machine learning algorithms[10]. Its principle is: through a given training data set, input new examples, find its nearest k examples in the data set, and then fit the model and instance classification[11]. This paper uses k-nearest neighbor to monitor the risks of cigarette business, and can analyze the keywords used in the labels extracted in the classification process, and then identify k related business risk problems, so as to carry out effective

supervision.

Calculating the distance between the points in the training dataset and the current point is usually done using the Euclidean distance formula:

$$dist(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)2} \tag{3}$$

3.2. Data Preparation and Label Setting

Table 1: Customer label table

number	label name	label rules	computational formula	statistical cycle	exclusionary rules
1	Large inventory	Taking the county company as the unit of calculation, the average quantity of the single specification cigarette inventory exceeds 200% of the average quantity of the specification cigarette inventory, which is determined as an abnormal label.	(Quantity of single specification inventory / average quantity of this specification inventory * 100%) is greater than or equal to 200%	month, week, date	Inventory is less than or equal to 1
2	Negative inventory	With the county company as the unit of calculation, the code scanning quantity of the single specification cigarette is greater than the cigarette inventory quantity of the specification, which is determined as abnormal label.	Single specification scan code outbound volume / the inventory of this specification is greater than 1	month, week, date	Excluding box sales
3	Cash accounts for a high proportion	With cash settlement, the amount of cash settlement accounts for a higher proportion of orders on the day.		month, week, date	Excluding box sales
4	Morning and evening single order scan code	6 to 9 am, 18 to 24 pm. Order multi-specification scan code sales.		month, week, date	Excluding box sales
5	Morning and evening more orders centralized scanning code	6 am to 9 am, 18 pm to 24 pm. orders more than centralized scanning code sales.		month, week, date	Excluding box sales
6	Holiday single order to scan the code	Single order multi-specification bar scan code sales.		month, week, date	Excluding box sales
7	Holiday many orders centralized scanning code	Multiple order bar centralized scan code sales.		month, week, date	Excluding box sales
8	Night single order scan code	0:00 to 6:00, single order multi-specification scan code sales.		month, week, date	Excluding box sales
9	Multiple orders at night are centralized scanning code	0:00 to 6:00, multiple orders centralized scanning code sales.		month, week, date	Excluding box sales
10	Single order scan code	During working hours, single order multi-specification bar scan code sales.		month, week, date	Excluding box sales
11	Multiple orders are centralized for scanning the code	During working hours, multiple order bar centralized scanning code sales.		month, week, date	Excluding box sales
12	Single order negative sales	Non-order cigarette order bar scan code into storage.		month, week, date	Excluding box sales
13	Frequent negative sales	Non-order cigarette multi-order scan code for storage.		month, week, date	Excluding box sales

We extracted the cloud pos data of the business system of each retail household, summarized the abnormal business behavior of cigarette retail customers into customer labels and regional labels,

and compiled the labeling rules, calculation formula, statistical cycle and exclusion rules according to the difficulty of abnormal data collection and processing, algorithm calculation power, regulatory requirements, etc.

Customer label: Customer labels are mainly formed in two ways, one is provided by the customer, namely explicit preference. This way is mainly provided through the information provided by retail customers during registration, such as household registration, business form, business circle and education; the other is inferred from the existing data, that is, during data processing. This way is mainly effective data found through data analysis, such as sales, inventory, brand type, etc. (Table 1).

Area label: a label that describes the common characteristics within a customer or group is called a common label, also called an area label. Regional label is a key element of group formation, equivalent to the greatest common divisor of a set of data, describing the attributes or behavioral tendencies shared by the data subject within a group (Table 2).

Table 2: Area label table

number	label name	label rules	computational formula	statistical cycle	exclusionary rules
1	Regional inventory decreased	Take the self-discipline group as the unit of calculation. Within one day, 3 households (including) and above the same specification cigarette inventory (strip) reduced, to be identified as abnormal label.	Single specification scan code output is more than or equal to 3	month, week, date	Excluding box sales
2	Regional inventory clearance	Take the self-discipline group as the unit of calculation. Within 1 day, 3 households (including) and above the same specification cigarette inventory (bar) cleared, determined as abnormal label.	Single specification scan code output / inventory of this specification is equal to 1	month, week, date	Excluding box sales
3	Centralized code scanning in the area	Within the region (in the same order cycle), more than 3 customers will sell the same specifications centrally.		month, week, date	Excluding box sales
4	After delivery, the area single specification inventory is 0	After ordering, the customer will scan the code centrally, and the inventory of more than 3 people in the area (in the same order cycle) is 0.		month, week, date	Excluding box sales
5	Regional single specification inventory decreased after delivery	After ordering, the customer conducts centralized scanning code, and the inventory of more than 3 people in the region (in the same order cycle) is reduced.		month, week, date	Excluding box sales
6	After delivery, the area multi-specification inventory is 0	After ordering, the customer will scan the code centrally, and the inventory of more than 3 people in the area (in the same ordering cycle) is 0.		month, week, date	Excluding box sales
7	Post-delivery area multi-specification inventory reduction	After ordering, customers conduct centralized scanning code, and the inventory of more than 3 people in the region (in the same order cycle) is reduced.		month, week, date	Excluding box sales

3.3 Establish the Model

This paper uses four model groups, naive Bayes algorithm, logistic regression and k-nearest neighbor. First of all, the missing values and abnormal values obtained in the cloud pos system are processed. Then, according to the different characteristics of customer labels and regional labels extracted from them, two dimensions of "commodity inventory" and "order sales" are selected to calculate the operating risk level and risk value, and average the final results. Secondly, 20 different

training sets and test sets are randomly selected through the downsampling method (selecting some data from a few data sets and recombining them into new data sets). In the model training and evaluation of indicators, the unbalanced data indicators are used to evaluate the model effects.

The model-building procedure is shown in Figure 4.

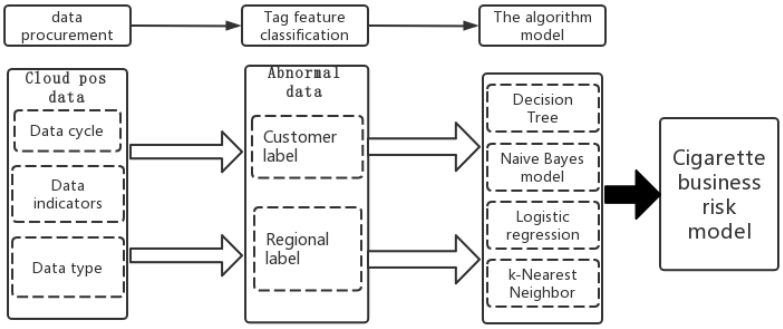

Figure 4: Cigarette business risk model

4. Research Conclusion

At present, in the practice of standardized management of cigarette operation, there are still risks in the processes of tobacco monopoly, marketing and internal management, which is not conducive to the stability and standardization of the tobacco market. Therefore, with the help of data mining technology, this paper uses different types of algorithms to build a tobacco business risk analysis model. On the basis of reasonable classification of various kinds of labels and indicators, the basic mode and characteristics of data-driven cigarette operation are sorted out, and the risk is effectively identified and evaluated according to the types of customer labels and regional labels. On this basis, it provides an effective reference for promoting the fine joint supervision of cigarette business activities and building a risk early warning system.

References

[1] Guo X. C. Application of a naive Bayesian classification algorithm. Communication World, 2019, 26 (01): 241-242.
[2] He X. N, Duan F. H. Linear regression case analysis based on the python. Microcomputer application, 2022, 38 (11): 35-37.
[3] Li C. S. Construction of enterprise financial risk early warning model based on logistic regression method. Statistics and Decision-making, 2018, 34 (06): 185-188.
[4] Li C. B. Based on the "risk-oriented" internal supervision mode of cigarette business. Chongqing and the World (academic edition), 2013.
[5] Luo X., Ouyang Y. X, Xiong Z., Yuan, M. The k-nearest neighbor-based collaborative filtering algorithm was optimized by similarity support. Journal of Computer Science, 2010, 33 (08): 1437-1445.
[6] Pi Y.C. Application of k-neighbor classification algorithm. Communication World, 2019, 26 (01): 286-287.
[7] Qing S. Construction and analysis of the company's business risk measurement method. Accounting and Communication, 2013 (24): 123-125.
[8] Shen M. Y, Han M., Du S. Y, Sun R., Zhang C. Y. Summary of integrated classification algorithms for data flow decision trees. Computer applications and software, 2022, 39 (09): 1-10.
[9] Xu Z. Q, Li Y. Y, Wan Y. C, Hu L. F, Xu B. Y. Thinking and measures on risk management theory. Value Engineering, 2020, 39 (05): 25-26.
[10] Zhang C. J. Decision tree algorithm for big data analysis. Computer Science, 2016, 43 (s1): 374-379 383.
[11] Zhang R., Wang Y. B. Machine learning and its algorithms and development research. Journal of Communication University of China (Natural Science edition), 2016, 23 (02).