Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 11
EMFT 2024

Research on Cigarette Market Capacity Forecasting Based on Data Mining Methods

Xudong Wu *

Anhui Tobacco Company Chizhou Company Information Center, Anhui, China, 247100

* Corresponding Author Email: tom2024267@gmail.com

Abstract. With the tobacco industry gradually stepping into the modern cigarette marketing mode of data and informationization, cigarette precision marketing will become the industry's new way of refined marketing management. Among them, precise placement is one of the important contents of cigarette precision marketing. Precision placement is based on the segmentation, quantification, and combination of brand, customer, market, and time, aiming at precisely supplying the brand to the corresponding market segments, and realizing the goal of “finding the market for the brand, finding the brand for the market; finding the brand for the customer, finding the customer for the brand”. Through the segmentation and quantification of the market and customers, this study this paper can more accurately understand the market demand and customer characteristics, to formulate more targeted marketing strategies. Accurate forecasting of cigarette market capacity is of great significance to enterprises and regulators, and the integration of multiple data processing techniques and forecasting models can help improve the accuracy and applicability of predicting. Specifically, this study aims to construct an accurate customer placement model based on multiple linear regression. After data analysis and model training, the prediction results of the model are validated and evaluated. The results show that the relative error between the model-predicted placement and the actual placement is 1.95%, and the model has high prediction accuracy and stability, which provides valuable reference information for the work of precise cigarette customer placement.

Keywords: Multiple linear regression; Principal component analysis; Brand placement; Precision placement.

1. Introduction

The cigarette market, as an important part of the consumer goods market, has a significant guiding significance for enterprises to formulate production and marketing strategies in terms of market capacity and demand forecasts[1]. Accurate forecasting of market capacity and sales demand can optimize resource allocation, improve economic efficiency, and provide support for government and industry regulation[2]. Traditional market capacity forecasting methods such as linear regression, although simple, have limited accuracy and applicability when facing complex markets influenced by multiple factors[3] [4]. With the development of data science and machine learning, researchers have improved the accuracy and reliability of market forecasting by incorporating techniques such as data mining, principal component analysis, multiple regression analysis, and neural networks [5].

The core of precision marketing is to clarify the target [6]. Therefore, in-depth research on the capacity, structure, and consumer trends of the cigarette market is needed to provide a basis for business decisions [7]. At the same time, this study this paper should improve the market monitoring and demand forecasting system, deeply analyze the characteristics of consumer groups, and identify the key points and difficulties in market development. The current market forecast mostly relies on historical data, which is difficult to accurately reflect changes in market demand and affects the effectiveness of marketing decisions [8] [9]. To solve this problem, this study adopts big data technology and multiple linear regression algorithms to comprehensively evaluate historical consumption data, regional population, economic level, and other indicators to construct a refined data model and improve prediction accuracy.

To address this, this study proposes to use advanced big data technology to build a refined data model. By adopting the multiple linear regression algorithm, the market capacity can be calculated synchronously [10]. After comprehensively evaluating a variety of relevant indicators, this study this paper selected the following series of data as inputs that are closely related to cigarette market capacity [11]: historical consumption data and consumption trends, historical cigarette consumption data, regional population size, regional CIP index, regional economic patterns, regional demographic structure, total size of the regional annual number of items, regional industrial index, regional number of retailers, and regional economic level. Considering that changes in market capacity are affected by a variety of factors, it is proposed to select ten indicators in three categories of cigarette consumption, cigarette sales, and economic development for multiple linear regression analysis, and combine them with Pearson correlation analysis to calculate the correlation coefficients of the ten indicators and make comparisons.

2. Data Cleaning and Regression Analysis

The data for this study were obtained from https://www.stats.gov.cn/, China Statistics Network, and other website data.

To ensure the learning and prediction effect of linear regression and neural networks, it is necessary to pre-process the original data.

2.1. Data Cleaning

Remove outliers and missing values. In the process of data collection, due to certain reasons, there may be missing values or abnormal values in the data. To ensure the accuracy and consistency of the data, these data need to be cleaned. In the case of sales data, no cigar cigarette sales data are counted, single-month market price outlier data - data below 80% of the monthly average market price - are deleted, and only sales data for the month after the introduction of new products are counted.

2.2. Data Standardization

Standardize the data of different indicators to the same scale. In this study, the min-max standardization method is used to standardize the data to between 0 and 1, which makes the weights between different indicators equal, and at the same time avoids the data affecting the learning effect of the model because of too large a difference in values.

3. Linear regression modeling

3.1. Linear regression modeling

Multiple regression analysis applies to situations where there is a dependence between multiple determining variables and one dependent variable, and it is difficult to determine the primary and secondary variables. In the study of the problem of market capacity forecasting, changes in market capacity are affected by multiple factors, so multiple regression analysis is needed.

Linear regression is a frequently used forecasting method, while linear regression models are often employed to describe the relationship between predictor variables and influencing factors in the study of specific problems with multiple factors. The linear model is characterized by its simple form and ease of modeling. At the same time, the theory also contains some important ideas in machine learning, and the coefficients before the variables in the established regression equations symbolize the importance of this influencing factor on the predictor variables, so the linear regression interpretability is very good.

Regression analysis for specific problems, if only one independent variable and one dependent variable are involved, and the relationship between the two variables can be approximated by a linear function, this type of regression is known as mono linear regression analysis; if two or more

independent variables are included, and the relationship between the two variables can be approximated as a linear relationship, then the regression is known as multivariate linear regression analysis.

3.2. Market capacity regression forecasting

In the study of the problem of market capacity forecasting, the change in market capacity is affected by a variety of factors, so it is necessary to carry out multiple regression analyses. After the team's preliminary sorting and analysis, this study this paper selected ten indicators in three categories: cigarette consumption, cigarette sales, and economic development.

Indicators in the category of economic development: the number of the regional population, regional GDP, and the number of regional retail households; the data content is from 2008-2022 for the city's four-county and district bureaus. The data sources are statistical yearbooks and government work reports of the CZ region and the districts and counties under its jurisdiction.

Cigarette sales indicators: average sales per household, sales volume, sales volume per carton; data content is cigarette sales data of the city's 4 county and district bureaus from 2008-2022; data source is the Tobacco Business System, Integrated Marketing Platform.

Cigarette consumption indicators: wholesale and retail gross margin, terminal year-end inventory, terminal year-end inventory stock-to-sale ratio, and source satisfaction; the data content is the data of 4 county and district bureaus in the city from 2008 to 2022; the data source is the data collection platform.

3.3. Normality test

For the sample data, this study this paper first carried out the normality test. From the histogram (Figure 1), it can be seen that the distribution of the data shows a bell-shaped curve, the highest point of the curve is located in the expected value of the normal distribution, and the left and right sides are roughly symmetrical, which indicates that the data obey the normal distribution. From the residual P-P plot (Figure 2), it can be seen that the scatter points are roughly distributed along a straight line, indicating that the distribution of the data approximately obeys the normal distribution.

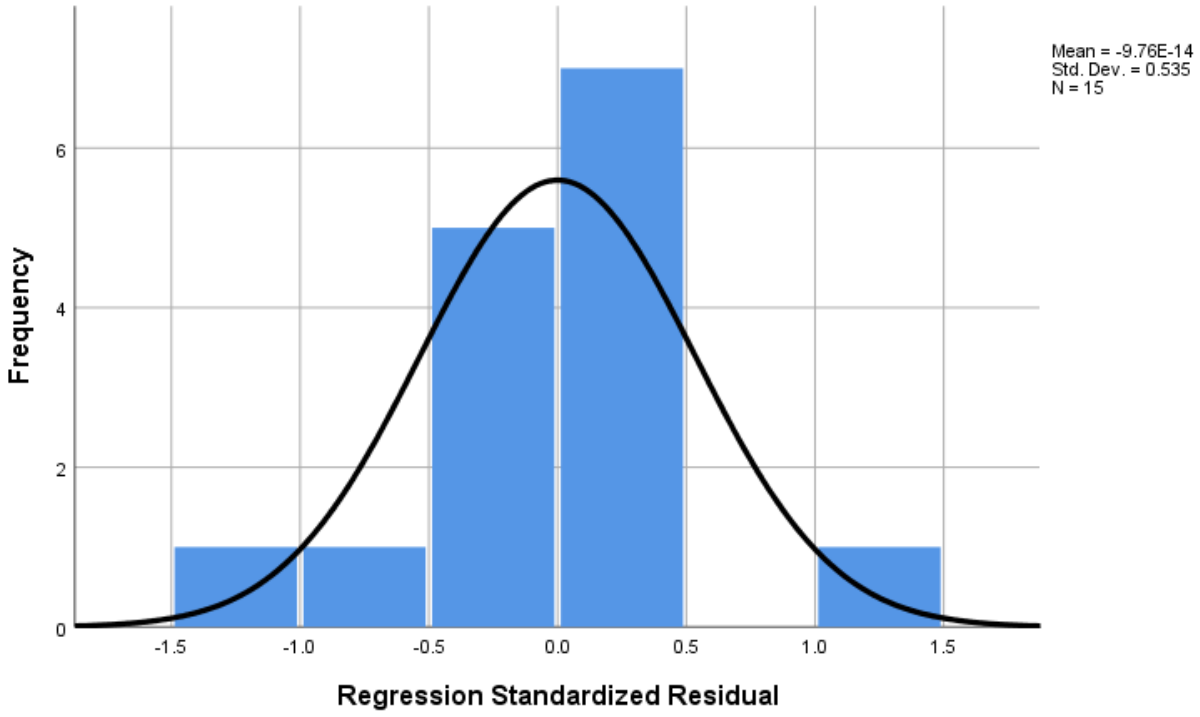

Figure 1. Histogram of normality test.

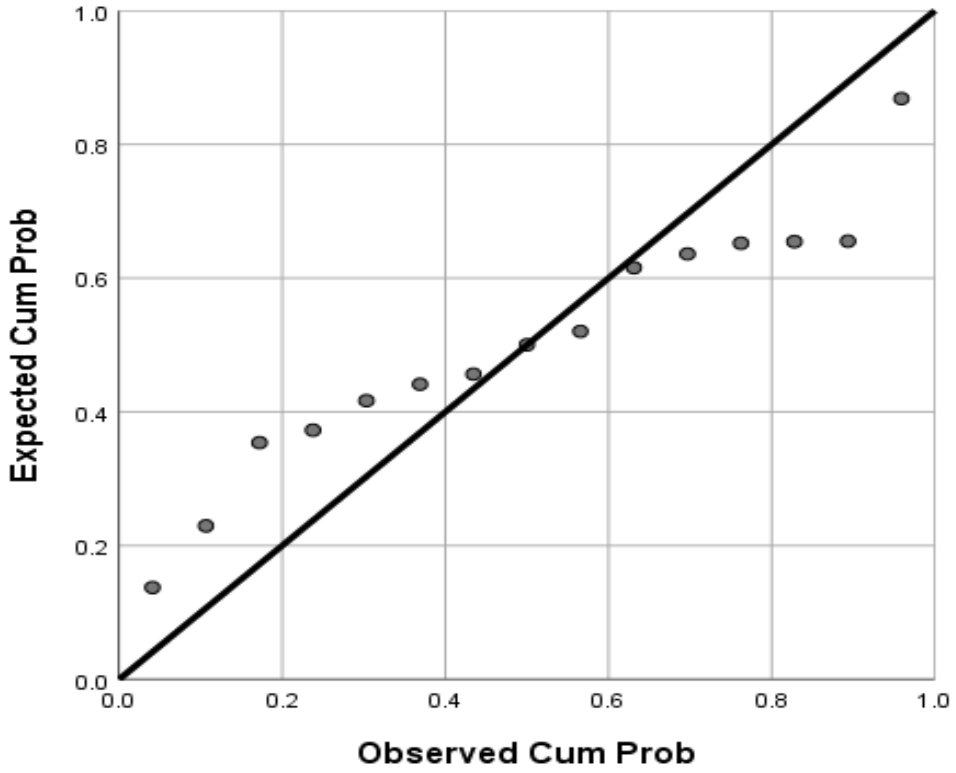

Figure 2. P-P plot for normality test.

3.4. Linear relationship analysis

Linear relationship analysis was conducted on the sample data, and the correlation coefficients of ten indicators were calculated and compared by Pearson correlation analysis, finally, 10 parameters with significant correlation were selected, including regional population size, regional GDP, regional number of retail households, average household sales volume, sales volume, sales volume per box, wholesale and retail gross margins, terminal end-of-year inventory, terminal end-of-year inventory stock-to-sales ratio, and satisfaction with the source of goods.

3.5. Sample independence analysis

For the sample data, this study this paper continue to carry out the independence analysis, using the Durbin-Watson method to calculate the test statistic D-W, measured by the software D-W value of 2.710, proving that it is basically in line with the sample independence, as shown in Table.1.

Table 1. Line Table of results of analytical tests of sample independence.

Model	R	R Square	Adjusted Square	Std. Error of theEstimate	Durbin-Watson
1	0.987[a]	0.975	0.912	765.413	2.710

The D-W values ranged from 0-4, with no (first-order) correlation as D-W approached 2. The model output was 2.710, which proved to be largely consistent with sample independence.

Table 1 gives a summary of the model, which has a compound correlation coefficient (R) of 0.987, a resolvability coefficient (R-squared) of 0.996, and a modified resolvability coefficient (adjusted R-squared) of 0.912, which indicates that the model's explanatory power is very good.

3.6. Modeling

Because of the linear relationship between the change in the amount of investment and several variables such as GDP, resident population, and total retail sales of social consumer goods, after the test of linearity, independence, normality, and chi-square, it meets the characteristics of the multiple linear regression model, and therefore the model is chosen for modeling.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n + \varepsilon \quad (1)$$

After using the standardized sample data set to set up a good regression model, the model coefficients obtained at this time are the standard regression coefficients, which can only reflect the relative importance of the variables, to better explain the role of the various influencing factors on the amount of delivery, it is necessary to convert the standardized regression coefficients into non-standardized coefficients. So the standard regression coefficients are reduced and then the non-standardized coefficients of each variable are derived, and the final expression of the non-standardized ordinary linear regression equation is as follows:

$$\begin{aligned} y = &-531124.470 + 0.653x_1 + 7.286x_2 - 6.898x_3 + 3265.471x_4 + 1376.232x_5 \\ &- 0.796x_6 - 1581.704x_7 + 11.948x_8 - 19860.410x_9 + 7326.824x_{10} + \varepsilon \end{aligned} \quad (2)$$

4. Elimination of Multicollinearity and Model Testing

4.1. Test for multicollinearity

Multicollinearity refers to the phenomenon of high correlation between independent variables in regression analysis, which leads to distortion of model estimation and a decrease in prediction accuracy. Commonly used methods for testing multicollinearity in SPSS include:

Correlation coefficient matrix:

First, the correlation coefficient matrix between all independent variables is calculated. If the correlation coefficient between two or more independent variables is high (usually considered to be greater than 0.8), multicollinearity may be a problem.

Variance inflation factor (VIF):

VIF is an important indicator for testing multicollinearity. In SPSS, the results of the regression analysis can be used to view the VIF values for each independent variable. If the VIF value of an independent variable is greater than 10 (the strict criterion is 5), then the variable is considered to have multicollinearity.

Introduced by Marquardt in 1906, the inverse of tolerance, when there is a covariance between independent variables, the variance of the regression coefficients estimated by the method of least squares increases by a multiple of the variance of the regression coefficients estimated when there is no covariance between the independent variables, and the greater the value of the VIF, the stronger the degree of multicollinearity between the variables. As with the correlation coefficient indicator of the independent variables, the critical value utilized to diagnose the problem of multicollinearity is not easy to determine. Some scholars suggest that when VIF $\geq$ 5 or VIF $\geq$ 10, it can be considered that there is a serious covariance between independent variables but the critical value will be different for different specific cases.

Tolerance:

Tolerance is the inverse of VIF (tolerance = 1/VIF), if the tolerance of an independent variable is less than 0.1 (the strict criterion is 0.2), the same variable is considered to have multicollinearity.

Norusis mentioned in 1982 that TOL=1-R2i, Ri is the coefficient of determination of a linear regression model obtained with the independent variable Xi as the dependent variable and the other variables as independent variables, and a small tolerance suggests the possible presence of covariance, and less than 0.1 indicates that the multicollinearity is serious.

According to Table 2, this model has the problem of multicollinearity.

4.2. Elimination of multicollinearity

In the study of the problem of market capacity forecasting, the change in market capacity is affected by a variety of factors, so it is necessary to carry out multiple regression analyses. After the team's preliminary sorting and analysis, this study this paper selected ten indicators in three categories: cigarette consumption, cigarette sales, and economic development.

Once the existence of multicollinearity is recognized, the following methods can be taken to eliminate or mitigate its effects:

Remove variables with high VIF:

If the VIF value of an independent variable is significantly higher than other independent variables, consider removing it from the model.

Combining highly correlated independent variables:

If there is a high correlation between multiple independent variables, try merging them into a new variable. This usually requires that these variables have theoretically similar meanings or roles.

Increasing the sample size:

Multicollinearity problems are sometimes associated with inadequate sample sizes. Increasing the sample size may help to mitigate the effects of multicollinearity.

Use regularization methods:

Regularization methods such as ridge regression and LASSO regression (L1 regularization) can be used to deal with multicollinearity in regression analysis by introducing penalty terms to limit the absolute value of the regression coefficients, thus improving the stability and accuracy of the model.

Perform factor analysis or principal component analysis:

Factor analysis or principal component analysis can reduce the effects of multicollinearity by transforming the original variables into new unrelated variables. These new variables (i.e., factors or principal components) are linear combinations of the original variables and are uncorrelated with each other.

5. Model Correction and Model Validation

5.1. Model Correction

Principal component analysis can reduce the effects of multicollinearity by transforming the original variables into new uncorrelated variables. These new variables (i.e. factors or principal components) are linear combinations of the original variables and are uncorrelated with each other.

The model further eliminates multicollinearity by applying principal component analysis. Performing Principal Component Analysis (PCA) is a commonly used data dimensionality reduction technique that simplifies complex datasets by extracting key features from the data.

KMO and Bartlett's test: the KMO value is 0.626>0.5 for comparison with 0.5; the statistical value of Bartlett's test of sphericity, the test p-value <0.05, sig<a, is suitable for principal component analysis, as shown in Table 2.

Table 2. KMO and Bartlett's test table.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.626
Bartlett's Test of Sphericity	Approx. Chi-Square	224.380
	df	45
	Sig.	0.000

Initial solutions for principal component analysis, as shown in Table.3.

Eigenvalue and variance contribution: Because the cumulative variance contribution of the first 2 principal components reaches 84.115% and the corresponding \lambda _[1] are 3.436 and 1.611 respectively, and all \lambda _[1] > 1, the first 2 principal components are selected.

Table 3. Table of initial solutions for principal component analysis.

	Initial	Extraction
Regional Population (in ten thousand)	1.000	0.432
Regional GDP (in 100 million yuan)	1.000	0.969
Number of Retail Outlets	1.000	0.947
Average Sales per Outlet (in boxes)	1.000	0.722
Sales (in 100 million yuan)	1.000	0.891
Sales per Box (in yuan) including tax	1.000	0.965
Wholesale and Retail Gross Profit Margin (%)	1.000	0.940
Terminal Inventory at the End of the Year (in boxes)	1.000	0.905
Inventory-to-Sales Ratio of Terminals at the End of the Year (in days)	1.000	0.949
Satisfaction with Goods Supply (%)	1.000	0.781

Gravel plot (Figure 3): after the 3rd eigenvalue, the trend of change begins to level off, so it is appropriate to take the first 2 principal components.

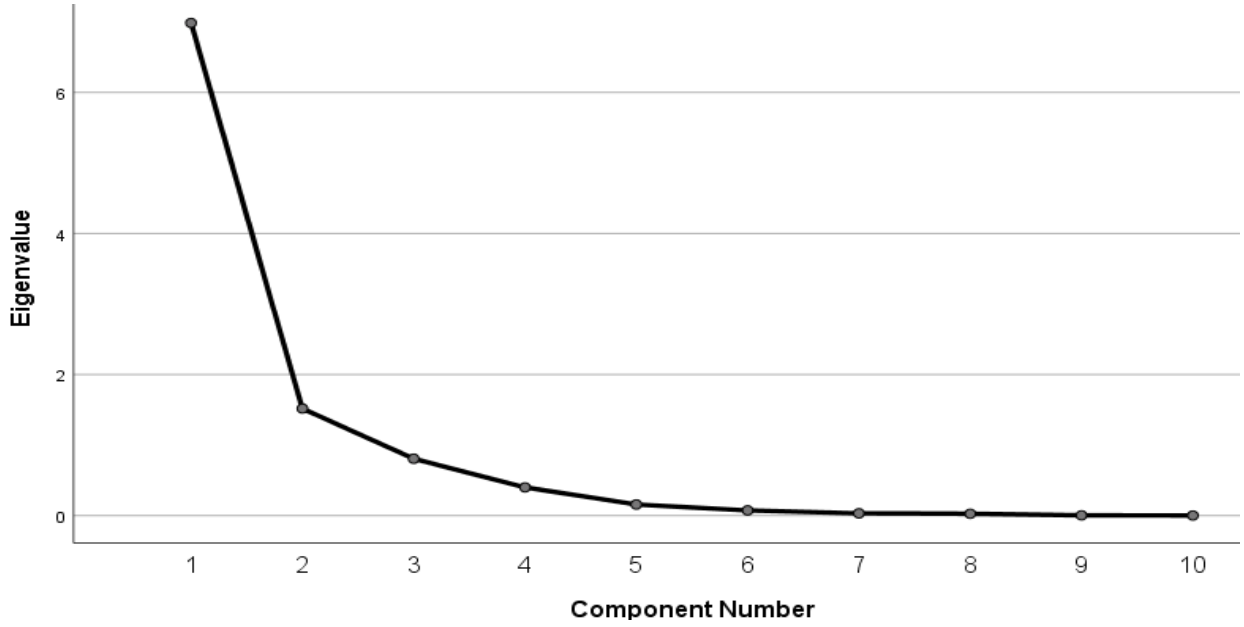

Figure 3. Gravel diagram.

Factor loading matrix: perceptions of development opportunities, perceptions of social status, attitudes toward job advancement, and leadership style preferences loaded highly on the 1st factor, so the 1st factor can be viewed as a composite variable of these variables.

2D simple scatterplot of principal components (if 2 principal components are extracted): insert X and Y axes, and label sample numbers. Samples falling into the first and fourth quadrants are relatively good and those falling into the second and third quadrants are relatively poor (Figure 4).

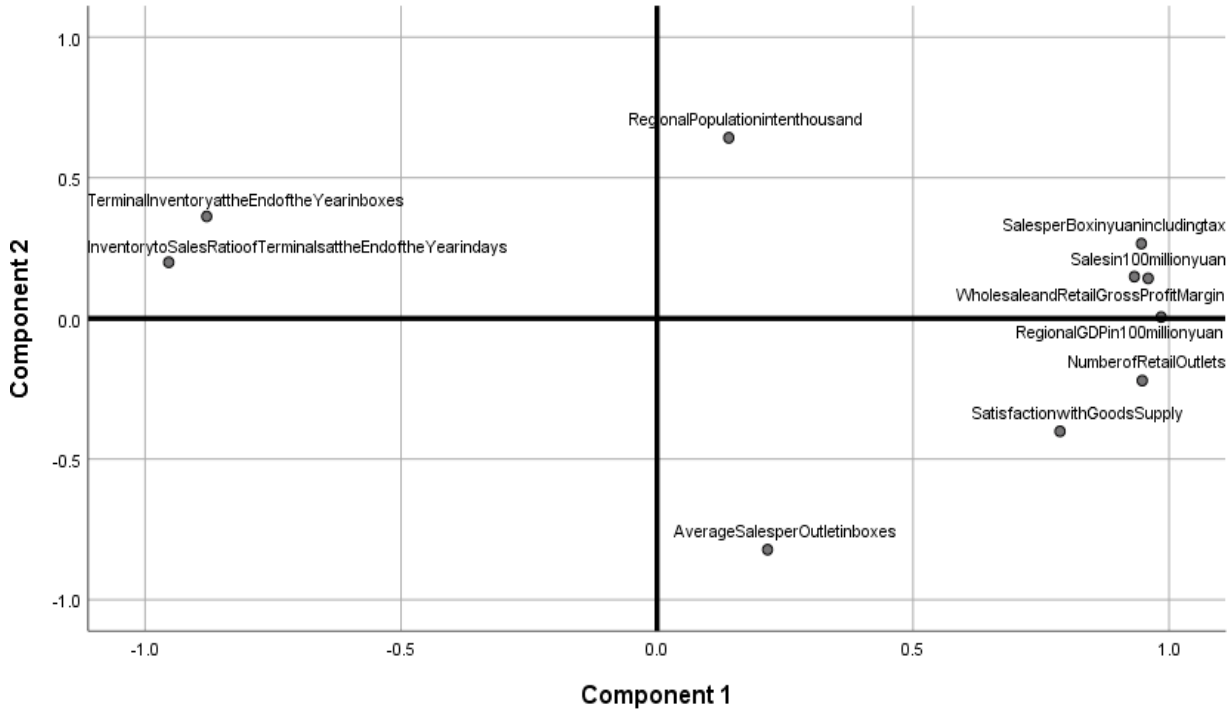

Figure 4. Two-dimensional simple scatterplot of principal components.

5.2. Model Validation

To prove the reliability and applicability of the model, the research team conducted several rounds of model testing. Firstly, the degree of fit analysis was carried out, and the results showed that the adjusted R-square value was 0.930, indicating that the independent variables of the newly established regression model could explain 93.0% of the reasons for the changes in the quantity put in place, as shown in Table.4.

Table 4. Table of model fitting results.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	0.992[a]	0.985	0.930	680.422	2.054

The test was conducted for the overall significance of the regression model, where F=17.956, p<0.001, proving that the model as a whole passed the F test. The results are shown below: as shown in Table.5.

Table.5. Table of F-test arithmetic results

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	91444820.982	11	8313165.544	17.956	0.018[b]
	Residual	1388923.951	3	462974.650		
	Total	92833744.933	14			

Through the above comparison and analysis results, it is found that the linear regression prediction error is small and the prediction accuracy is relatively high, so this study this paper use this model to estimate the cigarette launch quantity in 2023 based on the data of each variable index in 2022. The standardized data of each variable indicator in 2022 is inputted into the constructed cigarette quantity prediction model, and the approximate quantity of cigarettes to be put into the market in the coming year is calculated to be 52,415 cartons. To check the prediction effect of the model, the data of the company in 2022 is checked, and it is concluded that the quantity of cigarettes launched by the company in 2022 is 51,412 cases, and the relative error between the model's predicted quantity and the actual quantity launched is 1.95%.

Since the predicted results of the linear regression model are based on the assumptions that the state has not introduced any major policies for the cigarette market and the CZ market has not changed significantly, the predicted results have some limitations, but the model predictions can be used as a reference for the company.

6. Conclusions

This paper provides a research idea and framework applied to market forecasting-related areas, comprehensive analysis of the results of the modeling can be found that the linear regression prediction model has the advantage of higher prediction accuracy and relatively accurate prediction results. The linear regression model has a broader prospect in the field of sales volume forecasting which has more influencing factors, and the model can provide an effective auxiliary means for the company to make sales volume forecast estimation. With the help of linear regression as a data analysis method, the company can forecast future sales volume, and then the company can refer to the prediction results to formulate corresponding effective strategies.

References

[1] KSHETRI N. The evolution of the internet of things industry and market in China: An interplay of institutions, demands and supply [J]. Telecommunications Policy, 2017, 41(1): 49-67.

[2] VARMA V A, REKLAITIS G V, BLAU G E, et al. Enterprise-wide modeling & optimization—An overview of emerging research challenges and opportunities [J]. Computers & Chemical Engineering, 2007, 31(5): 692-711.

[3] FILDES R, MA S, KOLASSA S. Retail forecasting: Research and practice [J]. International Journal of Forecasting, 2022, 38(4): 1283-318.

[4] ARMSTRONG J S. Findings from evidence-based forecasting: Methods for reducing forecast error [J]. International Journal of Forecasting, 2006, 22(3): 583-98.

[5] KUMBURE M M, LOHRMANN C, LUUKKA P, et al. Machine learning techniques and data for stock market forecasting: A literature review [J]. Expert Systems with Applications, 2022, 197: 116659.

[6] YU C, ZHANG Z, LIN C, et al. Can data-driven precision marketing promote user ad clicks? Evidence from advertising in WeChat moments [J]. Industrial Marketing Management, 2020, 90: 481-92.

[7] KEKLIK S, GULTEKIN-KARAKAS D. Anti-tobacco control industry strategies in Turkey [J]. BMC Public Health, 2018, 18(1): 282.

[8] WITT S F, WITT C A. Forecasting tourism demand: A review of empirical research [J]. International Journal of Forecasting, 1995, 11(3): 447-75.

[9] SONG H, LI G. Tourism demand modelling and forecasting—a review of recent research [J]. Tourism Management, 2008, 29(2): 203-20.

[10] HONG J, WANG Z, CHEN W, et al. Online joint-prediction of multi-forward-step battery SOC using LSTM neural networks and multiple linear regression for real-world electric vehicles [J]. Journal of Energy Storage, 2020, 30: 101459.

[11] TI J, ZHENG Y, DUAN W, et al. Carbon footprint of tobacco production in China through Life-cycle-assessment: Regional compositions, spatiotemporal changes and driving factors [J]. Ecological Indicators, 2024, 165: 112216.