Analysis of customer satisfaction of HIMO Photo Studio based on Ordered Regresssion

Xinyi Sun[#], Zijie Wei[#], Jiaheng Wang[*, #]

[1] School of Statistics, Renmin University of China, Beijing, China, 100872

* Corresponding Author Email: loganlogan0807@163.com

[#]These authors contributed equally.

Abstract. With the upgrading of consumption and the improvement of living standards, a group of new type photo studios represented by HIMO Photo Studio have arisen and developed dramatically and become the benchmark of the era in the portrait photography industry. In this paper, we apply LDA topic classification and TF-IDF feature extraction method to conduct text feature extraction and encoding vectorization on the comment data of Beijing HIMO Photo Studio. An ordered regression model with customer satisfaction as the responding variable is established, and the interaction effect is also introduced into model to study the influencing factors of customer satisfaction creatively and quantitatively so that we can give suggestions for improving service quality. The empirical results show that factors such as photo season, shop type, waiting time, photo works, service attitude, photo subject, photo category and so on have a significant impact on customer satisfaction and there are significant differences in customer satisfaction with the same type of photos in different shops. Combining the business goals of HIMO Photo Studio and the industry needs of today's new type photo studios, this paper puts forward targeted suggestions for HIMO Photo Studio to further improve its services according to the conclusions of the model.

Keywords: HIMO Photo Studio, ordered regression model, LDA topic classification, TF-IDF feature extraction.

1. Introduction

Portrait photography has been a common form of entertainment and relaxation in nation's life. By 2020, there have been around 310,000 photography studios in China, employing about 4.25 million people with a CR4 of 0.059% and a CR8 of 0.086%, making it a competitive industry. In recent years we have witnessed the appearance of a large group of new type photo studio represented by " HIMO Photo Studio"(HIMO). These new type photo studio combines the advantages of traditional photo studios and art photo studios, providing both processive ID photo and simple art photo services, which gains extensive attention in the industry.

At the same time, HIMO as multiple shops has such problems as different work quality and inconsistent propagandization from shop to shop; its products have also been accused by customers of being distorted and not matching the original image. As new type photo studios adapt to the market and consumers, improving their services and increase customer satisfaction has become an essential and critical issue.

Sun Tianzhi [1] applys7Ps and other theories to study the current situation of service marketing of HIMO and proposes that HIMO should improve the moral quality of service personnel, personalize customer service, create a differentiated brand and other service marketing improvement measures to promote the competitiveness of HIMO. Wang Qidi [2] takes HIMO as the research object, establishes the value chain model of HIMO based on the value chain theory, analyzes its core value activities and discusses the elements of enterprise core competitiveness. Yu Yue et alt [3] analyzed the factors affecting the customer satisfaction of HIMO by holding a consumer survey and proposed that HIMO should focus on improving product quality and service level.

Previous studies are mostly based on marketing theories and questionnaire data so the research on customer satisfaction has certain limitations and subjectivity. In this paper, we propose the factors influencing customer satisfaction in hippocampus photo studios through feature extraction and

quantitative analysis of consumer comment data from HIMO in Beijing, and then give suggestions to improve service quality.

2. The basic theory of text feature extraction and ordered probit model

2.1. LDA Topic Model

In this paper, two text feature extraction methods are used: the LDA topic model and text classification based on TF-IDF metrics. The following part will focus on the basic theory of LDA topic model including model assumptions and parameter estimation.

LDA is a document topic analysis model based on the bag-of-words model, which assumes that every document has multiple topics, each of which corresponds to some different words. An article is constructed by first selecting a certain topic with a certain probability, and then selecting certain words under this topic with a certain probability. [4]

The process of estimating the parameters of the LDA subject model is relatively complex. As is shown in Figure 1, given hyperparameters α and β, based on the density function of the K-dimensional Dirichlet distribution with parameter α we can calculate the joint probability that a document belongs to the subject θ, the subject of each word z_j in this subject and each word w_j.[5]

$$p(\theta,z,w|\alpha,\beta)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)\,p(w_n|z_n,\beta) \tag{1}$$

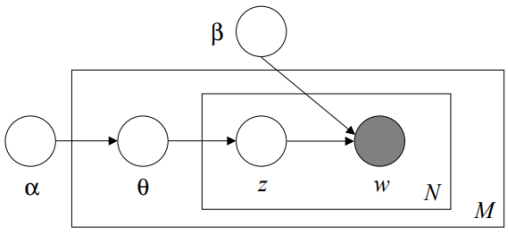

Figure 1. Schematic diagram of the model

After integrating the topics θ and z, we can obtain the marginal probability of a document.

$$p(w|\alpha,\beta)=\int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)\,p(w_n|z_n,\beta)\right)d\theta \tag{2}$$

By multiplying the probabilities of each document, we obtain the probability of the overall corpus.

$$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)\,p(w_{dn}|z_{dn},\beta)\right)d\theta_d \tag{3}$$

Based on this document generation principle and the word frequency before participle, we can conduct parameter estimation and inference of its topic based on Gibbs sampling and maximum likelihood estimation distribution.[6] Here we need to maximize a log-likelihood function of the form.

$$\ell(\alpha,\beta)=\sum_{d=1}^{M}\log p(w_d|\alpha,\beta) \tag{4}$$

The LDA model proposed by Blei in 2003 uses the E-M algorithm for maximization of the log-likelihood function.

2.2. TF-IDF metrics and text classification

The TF-IDF algorithm proposed by Salton[7] is the most used text feature extraction method. TF indicates how often a word appears in a document and IDF is the inverse document frequency, which

judges whether a word is important in the text. TF-IDF can show whether a word is important in the whole text and can distinguish different texts by doing the same for each text.[8] The TF and IDF indicators are calculated as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}},\ idf_i = \log^{\frac{N}{n+1}} \tag{5}$$

Where $n_{i,j}$ indicates the number of occurrences of the word w_i in a text d_j and the denominator of the tf formula indicates the number of words in d_j. N is the number of all texts contained in the data and n indicates the number of texts containing the word w_i in all datasets. The $tf-idf$ indicator for word w_i is the product of the two indicators above.

$$TF-IDF = tf_{i,j} \times idf_i \tag{6}$$

The $tf-idf$ indicator allows us to select top n important words in a document. After labelling each text, a decision tree model is constructed using the presence or absence of these words as indicators. By analyzing the node words and the purity of the nodes in the decision tree, the words that are decisive for text classification can be found.

2.3. ordered probit model

Regression analysis is a commonly used method for exploring and analyzing the relationships between variables. Within the framework of generalized linear model, the dependent variable in regression analysis can be binary variable, multinomial variable, count variable or a ordinal variable. Ordinal variables are classified into different categories of relative order (or have a natural ordering) at the measurement level. Ordinal variables are assumed to take numerical values to represent the order of a particular attribute, but these orders do not necessarily reflect the actual magnitude on a substantive scale, i.e. the distances between adjacent levels are not absolutely equal.[9] For example, consumer satisfaction after purchasing a product or experiencing a service can be categorized as unsatisfactory, fair and satisfactory, and there are relative degrees of highness and lowness between these variables, but clear and continuous distances between the different categories cannot be obtained empirically. In this regard, the order is generally reflected by assigning some integer to the above categories, with larger values indicating higher satisfaction.

Since the responding variable in ordered regression is discrete, we introduce the latent variable Z as a bridge to describe the relationship between explanatory variable X and responding variable Y.[10] Generally, we assume that Z is continuous so that we can construct an ordinary linear regression between the latent variable Z and X.

$$Z = \beta_0 + \beta X + \varepsilon \tag{7}$$

And for an ordered variable Y with k levels, we can set the thresholds $j = 1, 2 \ldots k-1$ to construct the relationship between Y and Z:

$$Y = \begin{cases} 1,\ Z < \alpha_1 \\ j,\ c_{j-1} \leqslant Z \leqslant c_j \\ k,\ c_{k-1} \leqslant Z \leqslant 1 \end{cases} \tag{8}$$

Just like treating the linear regression model, we make an assumption about the distribution of the random disturbance term ε in equation (7). The most common assumption is that it follows a standard normal distribution, and an ordered regression model based on this assumption is called a ordered Probit regression. Let $\alpha_j = c_j - \beta_0$, where the probability of $Y = j$ can then be expressed as:

$$\begin{aligned} P(Y=j) &= P(c_{j-1} \leqslant Z < c_j) \\ &= P(\alpha_{j-1} \leqslant \beta X + \varepsilon < \alpha_j) \\ &= P(\alpha_{j-1} - \beta X \leqslant \varepsilon < \alpha_j - \beta X) \\ &= F_\varepsilon(\alpha_j - \beta X) - F_\varepsilon(\alpha_{j-1} - \beta X) \end{aligned} \tag{9}$$

When assuming that ε obeys the normal distribution, we have $F_\varepsilon(x) = \Phi(x)$. Let the function $f_Y(X)$ denote the probability that for a given observation its responding variable Y belongs to the j class, the expression for $f_Y(X)$ can be known as follows:

$$f_Y(X) = \begin{cases} 1 - \Phi(\alpha_{k-1} - \beta X),\ Y = k \\ \Phi(\alpha_j - \beta X) - \Phi(\alpha_{j-1} - \beta X),\ Y = j \\ \Phi(\alpha_1 - \beta X),\ Y = 1 \end{cases} \tag{10}$$

Finally, maximum likelihood estimation is used to complete the estimation of the thresholds and the parameter β.

3. Model construction and model results

Through sampling design, the sample data includes comments in Dianping.com from 14 HIMO shops with three different types in Beijing and is consisted of a total of 18,995 comments. The structure of the obtained data is shown in table1.

Table 1. Variable specification table for comments data

Variable type	Variable name		Details	Range of values	Remarks
comments from Dianping.com	User ID		Character variables	/	Determined from crawl data
	comment stars		Qualitative variables (10 levels)	1star,1.5stars… 4stars, 4.5stars, 5stars	/
	Comment Level	satisfied	Character variables	/	Comments with 5 stars
		fair		/	Comments with 4+ stars
		unsatisfied		/	Comments with Under 4 stars
	Comment time		Character variables	27.08.2017 - 2022.10.19	Year/month/day
	Comment text		Character variables	/	Text data

Use LDA topic model and set the number of topics 5. All of 18,869 comments collected were analyzed to find the probability of each comment belonging to each topic and the keywords under each topic. The LDA results including topic name and keywords are shown in table 2

Table 2. LDA Theme Model Results

Topic number	**Topic name**	**keywords**
Topic1	Photo works	nice, service, effect, professional, environment, experience, efficiency
Topic2	Makeup and Photography	dressers, photographers, retouchers,0 aware, movement, atmosphere, emotion, adjustment, posture
Topic3	Service attitude	cute girl, gentle, patience, naughty, photography, retouching, makeup, careful, pretty
Topic4	Other	Children, staff, ceremony, tidiness, enthusiasm, attending school
Topic5	Waiting time	Time, hour, public, problem, queuing, combo, price, suggestions, pipelining

Combined with real situations, it was found that there was a close correspondence between the four subjects (waiting time, makeup and photography, service attitude and photo works) and the whole process of customers taking photos at HIMO. The probability of each comment belonging to these four subjects was therefore included in the model as an explanatory variable. In addition, to explore other factors which influence customer satisfaction, a text matrix for each comment was built using the TF-IDF text vectorization feature extraction method and a decision tree model was constructed with 3 comment level (satisfied, fair and unsatisfied) as classification criteria. Some nodes of the decision tree are shown in Figure 2.

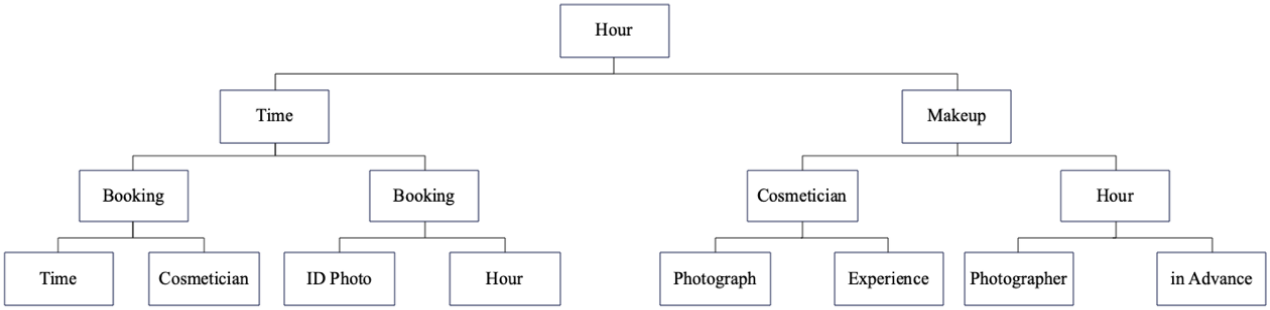

Figure 2. Text classification decision tree structure (excerpt)

The root node is word 'time', indicating that the most influential factor of consumer satisfaction is time. Based on the results of the text classification analysis, the three variables of waiting time, makeup and photography and type of photos should be included in the model as explanatory variable. In addition to the core variables above, three external variables and two variables related to photo attributes were added to the model. The final variable descriptions are demonstrated in table 3.

Table 3. Model variables

Variable type		Variable name	Detailed description	Range of values	Remarks
responding variable		comment level	Qualitative variables with three levels	satisfied, fair and unsatisfied	/
explanatory variable	External variables	Posted time	Qualitative variables with four levels	Q1, Q2, Q3 and Q4	/
		Shop Type	Qualitative variables with three levels	HIMO, HIMOMaster and HIMOFamily	/
		Media	Qualitative variables with six levels	Douyin,Xiaohongshu,Weibo, Dianping and live	/
	Internal variables	Waiting time	Quantitative variables	Integers from 0 to positive infinity	/
		Makeup and Photography			
		Service attitude			
		Waiting time			
		Photo works			
		Photo Subject	Qualitative variables with three levels	Self, parents and children	/
		Photo category	Qualitative variables, four levels	Personalized aesthetics, emotional documentation, ID photo and business photos	Official classification

3.1. Construction of the model

After data pretreatment as described in last section, we transform comment level, Photo Subject, Photo category, posted time, Shop Type and Media these five variables into factors to construct a ordered Probit regression model, and the variables were selected using stepwise regression according to the AIC criteria. Using the R4.2.2, the significant parameter estimates in Table 4 were obtained.

Table 4. Results of the regression model

Independent variable		Parameter estimates	t-value	p-value	Significance	Remarks
Quarterly	Q2	0.101	3.278	0.031	**	Baseline: Q1
	Q3	0.105	3.605	0.029	**	
	Q4	0.01	0.33	0.030	Not significant	
Shop Type	HIMOFamily	-1.409	-42.338	0.033	***	Baseline: HIMO
	HIMOMaster	-2.5	-43.595	0.057	***	
LDA Topic	Waiting time	-1.37	-28.881	0.047	***	
	Photo works	0.482	10.497	0.046	***	
	Service attitude	0.656	12.436	0.053	***	
Photo Subject	Child	-0.018	-0.381	0.046	Not significant	Baseline: yourself
	Parents	-0.318	-3.233	0.098	**	
Photo category	ID photo	0.171	6.4	0.027	***	Baseline: No category
	business photos	0.22	2.437	0.090	*	
	Emotional Records	0.241	5.348	0.045	***	
	Personalized aesthetics	0.246	3.602	0.068	**	
Threshold	1 | 2	-0.783	-21.229	0.037	***	
	2 | 3	-0.505	-13.712	0.037	***	

It can be seen that the stepwise regression method helped us delete the two variables of media and makeup and photography. Other variables in the model were all significant and AIC value of the model was 22433.76. The VIFs (variance inflation factor) for each variable are all less than 2, indicating that there was no multicollinearity between explanatory variables.

3.2. Exploring interaction

There are three types of HIMO being researched in data sample, and through actual research we have learned that each type of shop offers overlapping but not identical products to customers. For example, both the HIMOMaster and the HIMO offer ID photography, and both HIMO and HIMOFamily offer services such as wedding anniversary photography. In order to investigate whether the same type of photo is equally satisfied in different shops, we introduced a shop type-photo type interaction term into the model, and the results are shown in Table 5.

As we can see in table 5, the parameter estimates for the interaction term between HIMOMaster and ID photo are negative and significant, indicating that satisfaction with ID photos taken at HIMOMaster is significantly lower than at HIMO and HIMOFamily.

Table 5. Table of model results with the inclusion of interaction effects

Independent variable		Parameter estimates	SE	t-value	Significance	Remarks
Quarterly	Q2	0.102	0.031	3.289	**	Baseline: Q1
	Q3	0.106	0.029	3.623	***	
	Q4	0.01	0.03	0.349	Not significant	
Shop Type	HIMOFamily	-1.409	0.033	-42.34	***	Baseline: HIMO
	HIMOMaster	-2.427	0.061	-39.98	***	
LDA topic	Waiting time	-1.372	0.047	-28.9	***	
	Photo works	0.483	0.046	10.516	***	
	Service attitude	0.655	0.053	12.416	***	
Photo Subject	Child	-0.018	0.046	-0.381	Not significant	Baseline: yourself
	Parents	-0.316	0.098	-3.233	**	
Types of photographs	ID photo	0.187	0.028	6.584	***	Baseline: Other type
	Business photos	0.240	0.096	2.494	*	
	Emotional Records	0.232	0.050	4.611	***	
	Personalized aesthetics	0.259	0.074	3.523	**	
Interaction utility	HIMOMaster - Personalized aesthetics	-0.528	0.346	-1.527	Not significant	Baseline: HIMO
	HIMOMaster- ID Photo	-0.544	0.21	-2.586	*	
Threshold	1 | 2	-1.052	0.077	-13.7	***	
	2 | 3	-0.774	0.077	10.086	***	

4. Conclusions

Among the external factors influencing customer satisfaction in HIMO, the second and third quarters are significantly higher than the first and fourth quarters, so HIMO should reschedule its main annual campaign in the second and third quarters to produce better results. HIMOFamily and the HIMOMaster have significantly lower customer satisfaction than the HIMO, so HIMOMaster should conduct a consumer to ensure that the quality of its products is commensurate with the high prices. They could also try to expand their product range and adopt a more hands-on approach to marketing and product development, such as setting up shop experience days with discounts to increase customer goodwill.

In the service process, photo works and service attitude have a significant positive effect on customer satisfaction; waiting time has a significant negative effect on customer satisfaction; makeup and photography has no significant effect on customer satisfaction. Therefore, it can be seen that HIMO is very competitive in terms of photo effect and service attitude, which are the advantages that should maintain; waiting time, as the core issue affecting customer satisfaction, is the key aspect that HIMO should make changes to improve customer satisfaction. With regards for this, HIMO should optimize the workflow of each service, arrange for dedicated staff to coordinate time with customers, and set up an effective feedback mechanism for customer waiting time surveys to help companies control the relevant service standards in real time. Specifically, at the booking stage, the booking interval should be extended from the current twenty-minute interval to a larger half-hour interval, thus ensuring sufficient time to deploy subsequent make-up artists, photographers and other staff during the preparation phase before the photo shoot

In terms of photo subject, the satisfaction level of elderly customers at HIMO is low. As a representative brand of the new style photo studio, HIMO should maintain the business standard of the family photo studio while paying attention to the special service requirements of the elderly groups, such as preparing appropriate photo props for the elderly who have difficulty in moving around, adopting special make-up methods to reduce the damage of chemicals on the skin of the elderly, etc.; at the same time, it should also launch more photo themes or special activities based on the elderly groups to broaden the user group of HIMO. HIMO should also introduce more photo themes or special activities based on the elderly population to broaden the user base of the HIMO, in order to bring in more revenue and higher reputation. In terms of photo categories, customer satisfaction with ID photos is lower than with other types of photos. Some customers complained that the retouching was too "exaggerated" and that the photos were too different from the real image of the customer. In response to this technical problem, the technical staff at HIMO should take into account the feedback from customers on the results of their ID photos and adjust the retouching parameters, while the retouching technicians at the shops should also strengthen their communication with customers to ensure that the retouching results meet their requirements as much as possible.

References

[1] Sun Tianzi, Chang Ming. Analysis of hippocampus Photo studio service marketing [J]. Business Stories, 2020, (30): 22-24

[2] Wang Qidi. Analysis of Core Competitiveness of photography Enterprises based on value chain -- A case study of Hippocampus Photo Studio [J]. Business 2.0(Economic Management), 2021, (17): 100-101

[3] Yu Yue, Chen Meiling. Research on Customer satisfaction of hippocampus photo Studio [J]. Economic and Trade Practice, 2018, (24): 290,292

[4] Li Xinning. Stock price trend forecast based on text theme identification research [D]. Shandong business school, 2022. The DOI: 10.27903 /, dc nki. GSDSG. 2022.000042.

[5] David M. Blei; Andrew Y. Ng; Michael Jordan.Latent Dirichlet allocation (Article) [J]. Journal of Machine Learning Research, 2003, Vol.3: 993-1022

[6] Zhang Teri, Chen Yu. Factors extracted based on the LDA model health information users transfer behavior research [J]. Journal of library intelligence, 2019 tobacco (21): 66-77. The DOI: 10.13266 / j.i SSN. 0252-3116.2019.21.007.

[7] Salton, G.; Buckley, C.. Term-Weighting Approaches in Automatic Text Retrieval [J]. Information Processing and Management: an International Journal, 1988, Vol.24(5): 513-523

[8] Dong Jing. Discovery of hot topics on the Internet Based on Topic model and clustering algorithm [D]. Hebei University,2019.

[9] By ANNA A. O'CONNELL; Translated by Zhao Liangyun; Wu Xiaogang (Ed.). logistic Regression Model of Sequential Dependent Variables [M]. Gezhi Publishing House, Shanghai People's Publishing House, 2012

[10] Fang Xiaoying. The influence factors of tourism platform product sales analysis and forecast [D]. The northeast university of finance and economics, 2022. The DOI: 10.27006 /, dc nki. Gdbcu. 2022.001178.