J Inf Process Syst, Vol.19, No.4, pp.498~512, August 2023
https://doi.org/10.3745/JIPS.04.0282

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Factors Affecting the Sales of Newspapers and Magazines Based on Concise Catalog

Dayou Jiang*

Abstract

The traditional newspaper industry faces the opportunities and challenges of industry transformation and integration with new media. Consequently, the catalogs of newspapers and magazines are also updated. In this study, necessary information on catalogs was obtained and used to analyze the overall development trend of the newspaper industry. A word frequency analysis was then performed on the introduction and product categories of the catalogs, and the content and types of newspapers and magazines were examined. Furthermore, related factors such as price, number of pages, publishing frequency, and best-selling status were analyzed; the correlation among factors affecting best-selling status was also explored. Subsequently, each element and a combination of elements were used to generate a dataset, build three classification models, and analyze the accuracy of predictions of whether newspapers sold well under other circumstances. The experimental results showed that price is the most critical factor affecting the best-selling status of newspapers and magazines. Publishing frequency and the number of pages were also found to be significant indicators that impact people's subscription choices. Finally, a competitive strategy regarding content, price, quality, and positioning was developed.

Keywords

Classification, Competitive Strategy, Correlation Analysis, Descriptive Statistics, Magazines, Newspapers, Prediction

1. Introduction

The postal catalog of newspapers and magazines reflects the publication status of each year. The catalog is a bibliographic overview of various newspapers and magazines published throughout the country. It plays a vital role in regulating the publication, distribution, and publicity of newspapers and magazines. As newspapers and magazines have been transformed and reformed, today's catalogs contain more information. The concise catalog [1] contains substantial amounts of information, including postal code, name, publishing office, issue type, subscription type, the minimum number of consecutive issues, retail unit price, subscription unit price, monthly price, yearly price, issuing mode, layout, number of openings, introduction, postal signs, product categories, best-selling status, bundle specifications, distribution restrictions, remarks, and approval date.

The existing literature on condensed catalogs focuses on two aspects: ordering strategy [2,3] and

Manuscript received April 22, 2022; first revision June 16, 2022; second revision July 26, 2022; accepted August 10, 2022.
* **Corresponding Author:** Dayou Jiang (ybdxgxy13529@163.com)
Dept. of Computer Science, Anhui University of Finance and Economics, Bengbu, China (ybdxgxy13529@163.com)

pricing strategy [4-6] for specific types of newspapers and magazines. The shortcoming of the current research is that it stops at the descriptive statistical analysis of prices and recommends pricing strategies from the marketization perspective while ignoring the internal factors of newspaper and magazine popularity. This paper aims to analyze the critical information in the concise catalog of newspapers and magazines, explore the factors influencing the best-selling status attainment, and use the relevant influencing factors to predict popularity. The findings would benefit the newspaper industry and enhance market competitiveness.

A flowchart of the contents of this paper is shown in Fig. 1.

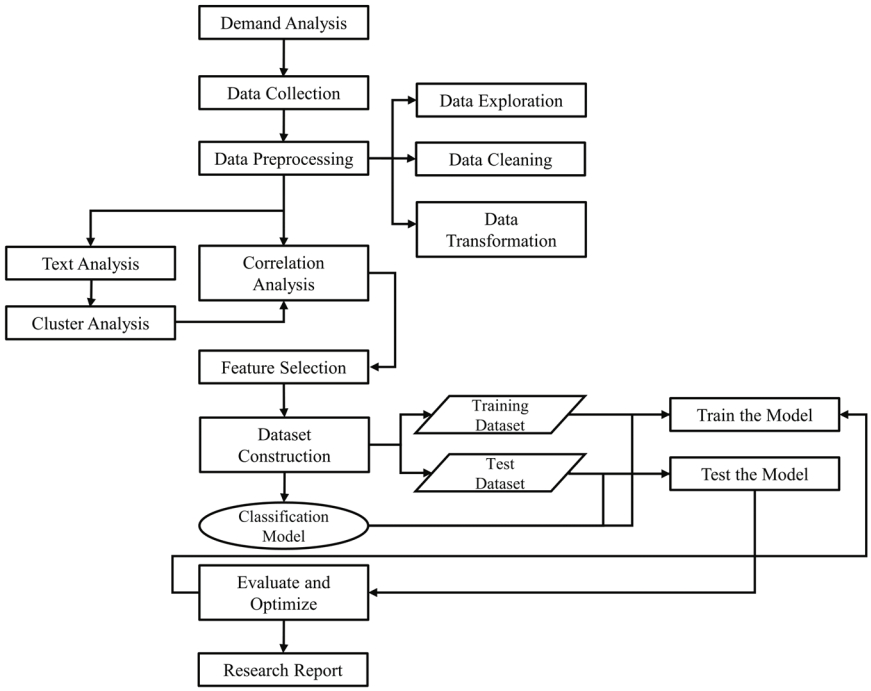

Fig. 1. Flowchart of research content.

As illustrated in Fig. 1, the rest of this paper is arranged as follows. Section 2 discusses the development trends of newspapers and magazines. Section 3 describes data exploration, data cleaning, data transformation, text analysis, and feature construction through correlation and clustering analyses. Section 4 describes the use of classification models to explore factors that influence sales. Section 5 presents several competitive strategies for newspapers and magazines for the media. Finally, Section 6 presents the conclusions and limitations of the study.

2. Background

2.1 Current Status

The development of China's print media has been declining recently, and the prospects of print media are pessimistic [7]. New media such as the Internet, television, and mobile newspapers can quickly

deliver information to a larger audience. Hence, their strong development momentum has placed considerable pressure on the survival of traditional newspapers. Problems such as the declining number of readers and rising costs have forced more publishing firms to change business strategies, for example, by compressing publication print volumes and increasingly focusing on new media.

Since 2005, newspaper and magazine distribution has experienced certain adjustments and fluctuations. The market value of newspapers has been declining, and many newspapers now need government intervention to survive [8]. In the current transition period, newspapers and magazines have been operated as a group while the capital scale of publishing, printing, and distribution groups is gradually increasing. By the end of 2017, 125 publishing and media groups were in China, including 47 newspaper groups [9]. Currently, capital is flowing into the media market at an accelerated rate while media houses seek to go public. Meanwhile, other scenarios exist, such as listings, mergers and acquisitions, and restructuring. Newspapers and magazines are subdivided into regions, industries, and groups of people. To go global, foreign media enter the local market through copyright cooperation and domestic means.

2.2 Development Trend

According to the “News and Publication Industry Analysis Report” [10] released by the China Press, Publication, Radio, and Television Network, the government has financed the development of the newspaper industry since 2016 to consolidate the critical position and role of mainstream newspapers as central in dispensing news and shaping public opinion [11]. In 2019, the decline in the total print volume of newspapers and magazines was halted, and profits began increasing. The average selling price of newspapers and magazines has increased while the average number of prints has decreased.

China's media integration has entered an era of transformation from a formal “merge” to an all-around “integration.” As a benchmark for the change and integration of the newspaper industry, the “two micro and one end” (Weibo, WeChat, client) platform has advanced significantly and become the core force of the new media in the newspaper industry. However, the print run of various newspapers has continued to shrink. For example, Life Services and Digest has declined strongly, while Professional Digest and Reader's Digest have declined slightly. The number of prints in various magazines has also declined, with the most apparent decrease occurring in literature and art. By contrast, the reduction in culture and education publications has eased. From the perspective of the content structure of newspapers, the proportion of comprehensive newspapers, life service newspapers, and abstract newspapers has continued to decline—the proportion of professional newspapers and magazines, such as teaching assistant newspapers, has increased. Additionally, the percentage of newspapers targeting older people and women has also increased. From the perspective of the product structure of magazines, the proportion of philosophy and social sciences, culture, and education has continued to decline, and the ratio of comprehensive magazines has increased moderately. Finally, the share of nature with arts, natural sciences, and technology has continued to decline.

3. Data Analysis

This section presents a descriptive analysis of critical information from concise catalogs and data preprocessing operations for feature selection and construction.

3.1 Data Preprocessing

The original 2021 postal catalog contains 1,478 newspapers and 7,415 magazines. As not all of the information is directly accessible, some of the information requires pre-processing, such as removing outliers, filling missing values, and converting data types.

3.1.1 Outlier removal

An outlier is a single value in a sample that deviates considerably from the rest of the observations in the model to which it belongs. The empirical rule states that 99.7% of normally distributed data lies within three standard deviations of the mean. An outlier is a measured value that differs from the mean by more than three standard deviations. The purpose of a study typically determines how to set the outliers to be removed from the data because outliers may introduce significant bias to the results of the analysis [12].

The primary distribution of the subscription unit prices of newspapers and magazines is presented in Table 1.

Table 1. Statistical summary of unit prices of newspapers and magazines

	Count	Mean	Median	Std. dev.	Min	Max	Mode
Newspapers	1,478	2.06	1.5	2.19	0.1	25	1.5
Magazines	7,415	26.36	16	44.33	2	1,520	10

The maximum value deviated considerably from the median and mode values, as specific types were not excluded. Therefore, the annual magazines and monthly newspapers were removed first. The subsequent statistical summary of the unit prices of newspapers and magazines is presented in Table 2.

Table 2. Statistical summary of unit prices of newspapers and magazines after removing specific types

	Count	Mean	Median	Std. dev.	Min	Max	25%	75%
Newspapers	1,237	1.78	1.5	1.39	0.1	20	1.05	2
Magazines	6,707	24.73	16	30.42	2	400	10	28

According to the empirical rule, the highest value for newspapers was $1.78 + 3 \times 1.39 = 5.95$, and the highest value for magazines was $24.73 + 3 \times 30.42 = 116$. Compared with Table 1, Table 3 shows that 273 newspapers and 832 magazines were removed as outliers.

Table 3. Statistical summary of unit prices of filtered newspapers and magazines

	Count	Mean	Median	Std. dev.	Min	Max	Mode
Newspapers	1,205	1.61	1.5	0.83	0.1	5.8	1.5
Magazines	6,583	21.49	16	16.49	2	112	10

3.1.2 Data transformation

We created numerical circulation indicators based on the categories. For example, the months were converted to a range of 1 to 12, the weeks to 1 to 52, and the days to 1 to 365.

3.1.3 Handling missing values

The missing values were handled by removing rows or columns with null values. However, there were 829 (187 newspapers + 642 magazines) blanks in the introduction of newspapers and magazines. Thus, another approach was to fill the blanks with the corresponding names.

3.1.4 Selection of best-selling indicators

The best-selling newspapers and magazines issued by China Post were selected by the Newspapers and Magazines Bureau of China Post Group Corporation from the tens of thousands of newspapers and magazines published and distributed nationwide. They were determined by calculating scores according to the number of recommendations, issue volume, year-on-year increase, and turnover. The selections were divided into six categories: current politics and finance, cultural integration, elderly health, family life, youth education, and fashion.

3.2 Descriptive Statistic

Distribution characteristics mainly analyze the frequency distribution, central type tendency, discrete trend skewness, and kurtosis. After removing outliers, 6,583 magazines and 1,205 newspapers were left. Newspapers and magazines differ considerably in price; hence, both categories were analyzed separately. The frequency distribution of newspapers is listed below:

{'Daily':109, 'Six Times Weekly':36, 'Five Times Weekly':113, 'Four Times Weekly':23, 'Thrice Weekly':68, 'Twice Weekly':123, 'Twice Ten Days':1, 'Weekly':704, 'Semi-Monthly':28}

The frequency distribution of magazines is listed below:

{'Twice Weekly':11, 'Weekly':35, 'XunKan':84, 'Twice Monthly':379, 'Monthly':3447, 'Bimonthly':2010, 'Quarterly':456, 'Semi-Annual':161}

The publication frequency of newspapers was mainly weekly, and magazines were mainly monthly and bimonthly. A descriptive statistical analysis was performed on newspapers and magazines; the results are presented in Tables 4 and 5.

Table 4. Descriptive statistics of newspapers

	Mode	**Mean**	**Median**	**Std. dev.**	**Min**	**Max**	**Skewness**	**Kurtosis**
Unit price	1.5	1.61	1.5	0.83	0.1	5.8	1.76	3.90
Yearly price	72	157.65	100	126.18	10.44	720	1.47	1.73
Number of pages	4	9.70	8	9.87	12	88	3.49	16.54
Publishing frequency	52	121.05	52	107.48	24	365	1.27	0.09

Table 5. Descriptive statistics of magazines

	Mode	Mean	Median	Std. dev.	Min	Max	Skewness	Kurtosis
Unit price	10	21.49	16	16.49	2	112	2.26	6.38
Yearly price	120	216.56	150	209.67	15	2,400	3.24	16.19
Number of pages	80	116.31	100	65.34	4	1,920	4.68	91.96
Publishing frequency	12	10.60	12	6.36	2	52	2.598	11.28

Table 4 shows that the dispersion of the unit price was small; the yearly cost and publishing frequency exhibited a platykurtic distribution, while the unit price and the number of pages exhibited a leptokurtic distribution, all of which were right-skewed-distributed.

Table 5 shows that the publishing frequency dispersion of magazines was minor; they were all leptokurtic-distributed. The publishing frequency was left-skewed, and the rest were right-skewed.

3.3 Word Count and Keyword Statistics based on Brief Introduction and Product Categories

The newspaper and magazine types were grouped from different angles. The word count analysis of the "Introduction" and "product category" information approximately classified their types. The "Introduction" uses different words and expressions to describe various newspapers and magazines briefly. For example, the "People's Daily" is described as "the organ of the Central Committee of the Communist Party of China." The description of "China TV News" is "announcement and promotion of CCTV programs, while taking into account the satellite programs of various provincial stations, reporting the situation on the screen and behind the scenes and the latest developments in the film and television industry at home and abroad, and providing extended services related to programs."

The top 15 keywords from the brief introduction of newspapers are listed below. In addition to news reports, newspapers currently focus on student-related content, and there has been a recent trend of an expanded distribution of these types.

{'Report':376, 'Training':350, 'Student':323, 'Learning':258, 'News':161, 'Service':134, 'Information':133, 'Knowledge':122, ' Sync':114, 'Ability':91, 'Method':82, 'Textbook':82, 'New':80, 'Economy':80, 'School Students':80}

The top 15 keywords from the brief introduction of magazines are listed below. The content of the magazines still mainly reflects domestic and international achievements as well as new trends in science and technology.

{'Research':2339, 'New':1949, 'Report':1640, 'Technology':1335, 'Theory':1175, 'Development':980, 'Field':963, 'Education':927, 'Domestic and Foreign':815, 'Science and Ttechnology':685, 'Management':640, 'Achievements':620, 'Application':556, 'Science':528, 'Clinical':518}

The Product Category descriptive words are not selected from the product structure of the hierarchical and content structure of magazines or newspapers. Moreover, the product category descriptions are not detailed. Therefore, overlap and confusion are inevitable in the classes. For example, university magazines are divided into finance, science and technology, agriculture, forestry, animal husbandry, medicine, philosophy, and natural sciences. Similarly, finance and economics have many subdivisions, such as finance, accounting, statistics, banking, securities, and insurance.

The top 15 keywords from the product category of newspapers are listed below. The most-occurred had an occurrence rate of 407/1,205 = 0.3378. The proportion of newspapers in market segments such as teenagers, senior citizen and women was high.

{'Teaching Aid':407, 'Infants':130, 'School Students':130, 'Children':130, 'Organ News':114, 'Comprehensive':61, 'Economy':54, 'Management':52, 'Education':49, 'Teaching':49, 'City News':45, 'Morning News':45, 'Evening News':45, 'Elderly':43, 'Women':43}

The top 15 keywords from the magazine product category are listed below. The highest-frequency word "Education" appeared in 1,038 magazines. Compared with the product categories of newspapers, those of magazines have more subdivisions, with education and colleges accounting for a large proportion. Medical and health research is gradually advancing, whereas geography and politics research is relatively rare.

{'Education':1038, 'Primary and Middle School':870, 'Medicine':867, 'Health':694, 'General':554, 'University':522, 'Journal':522, 'Technology':365, 'Social Science':306, 'General':250, 'Science':227, 'Economy':195, 'Management':192, 'Geography':179, 'Politics':158}

3.4 Feature Selection and Generation

This section describes the steps involved in feature construction and selection. First, two new features, product category, and profile category, were constructed through word frequency and cluster analyses; then, the impact factors related to best-selling were selected through correlation analysis.

3.4.1 Word frequency matrix

In the experiment, word segmentation was performed on each newspaper and magazine, and the text type data were converted into a numerical type to construct feature variables. The “jieba” (结巴) library [13] for Chinese word segmentation and the function CountVectorizer() were used to vectorize the text. Each word obtained after text vectorization represented a different numbered feature. The resultant database still required a manual screening of the words that needed to be eliminated. Certain phrases with similar meanings were also merged. Keywords in the introduction generated 3,140 words from newspapers and 11,380 words from magazines, whereas keywords in the product category generated 71 words from newspapers and 180 words from magazines.

3.4.2 Cluster analysis

After building the word frequency matrix, each newspaper and magazine was converted into a containing only 0 and 1 second. Next, the K-means clustering algorithm was used for cluster analysis. The K-means algorithm divides N instances into K disjoint "sample clusters" so that each point belongs to the cluster corresponding to its nearest mean. The algorithm reassigns objects to the clusters closest to the updated standard through an iterative relocation process within a close spatial range [14]. K-means clustering can efficiently cluster data and is often used as a pre-processing step for other algorithms.

Newspapers were divided into five categories according to their content: comprehensive, professional, life services, readers, and abstracts. Magazines were divided into five categories from the product structure: philosophy–social sciences, culture–education, literature–arts, natural sciences–technology, and general—the number of types K was set to 5. Here, the t-distributed neighborhood embedding nonlinear dimensionality reduction algorithm [15] was used to reduce the dimensionality of the data before implementing the K-means algorithm to ensure that the points in the generated clusters were relatively concentrated.

The categories can also be manually categorized according to the content structure of newspapers and the product structure of magazines. As the process is cumbersome and lacks clear standards, automatic clustering methods were used in this study.

3.4.3 Correlation Analysis

After removing outliers, 181 best-selling magazines and 51 best-selling newspapers remained. A high correlation was found to exist between the best-selling status and unit price, yearly price, the number of pages, and publishing frequency. Figs. 2 and 3 illustrate the results of the correlation analysis using the Pearson correlation coefficient [16] to measure the information from newspapers and magazines, respectively.

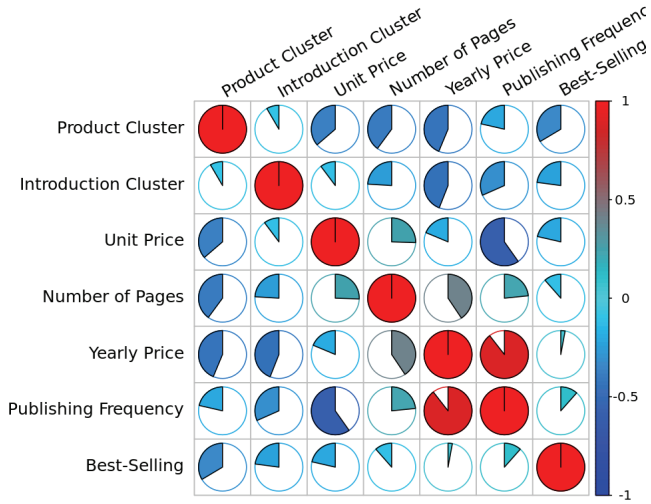

Fig. 2. Correlation between factors of newspapers.

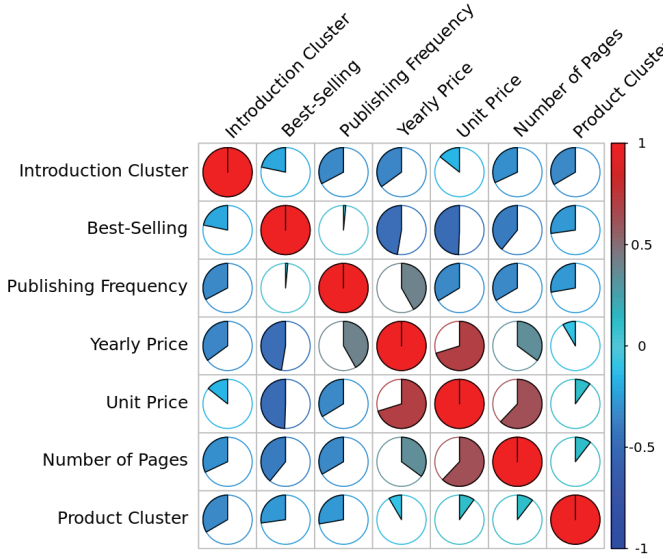

Fig. 3. Correlation between factors of magazines.

The popularity of newspapers was positively correlated with the yearly price, publishing frequency, and the number of pages, and the popularity had a weak negative correlation with the product cluster. The unit price was positively correlated with the yearly price and number of pages but negatively correlated with the publishing frequency.

For magazines, best-selling status was positively related to publishing frequency. The unit price, the number of pages, and the introduction cluster correlated more with best-selling status than the yearly price and product cluster. The unit price was positively correlated with the yearly price and the number of pages but negatively correlated with publishing frequency.

3.4.4 Feature selection

In addition to correlation analysis, we chose the unit price, publishing frequency, number of pages, and product clustering as the influencing factors of whether a newspaper sells well. Similarly, we chose the annual price, publishing frequency, number of pages, and introduction clustering as the influencing factors of whether a journal sells well. The correlation between the two clustering categories and whether they were famous was low. On the one hand, the automatic clustering method was defective; on the other hand, the introduction of the relevant text was too concise and could not contain too much information. Regardless, both were selected as the candidate factors.

4. Classification and Prediction

This section describes the building and training of a machine learning model for predicting whether newspapers and magazines sell well.

4.1 Data Processing and Model Building

4.1.1 Feature and target variables extraction

The target variable was set to “sell well” or “not,” and the independent feature variables were yearly price, publishing frequency, and product category.

4.1.2 Training and test sets division

The sample proportion of the target variable was highly unbalanced. The ratio of best-selling magazines was 181/6,583 = 2.75%, and the percentage of best-selling newspapers was 51/1,205 = 4.23%. In this case, it was necessary to construct the dataset by adopting the minority class sample oversampling or the majority class sample undersampling method. The synthetic minority oversampling technique (SMOTE) [17] was used to analyze the best-selling samples of the minority. The technique artificially synthesized new models according to the best-selling pieces to augment the dataset. Undersampling divided the majority class samples into equal parts with the number of instances of the minority class and formed the dataset in batches with the minority samples. The size of the dataset constructed by the undersampling method was limited, significantly impacting the experimental results. Hence, the SMOTE oversampling method was selected for the investigation. Owing to the limited sample, the k-fold cross-validation was used to evaluate the machine learning model. The dataset was shuffled randomly and split into five groups.

4.1.3 Classification Models Building

Three classification machine learning methods were used: support vector machine (SVM) [18], multilayer perceptron (MLP) neural network [19], and decision trees (DTs) [20] for model building.

When a training set is extensive or has many features, the linear kernel function (SVM without kernel function) is often used. MLP is a multilayer feedforward network model composed of multiple node layers. The MLP neural network overcomes the inability of the perceptron to identify inseparable linear data, can continuously optimize the network model through adaptive learning, and has strong computing power. DTs are constructed using a heuristic method called recursive partitioning.

4.2 Predicted Classification

The results presented below are the average results of 20 training and testing runs. The prediction accuracies under different factors were tested separately.

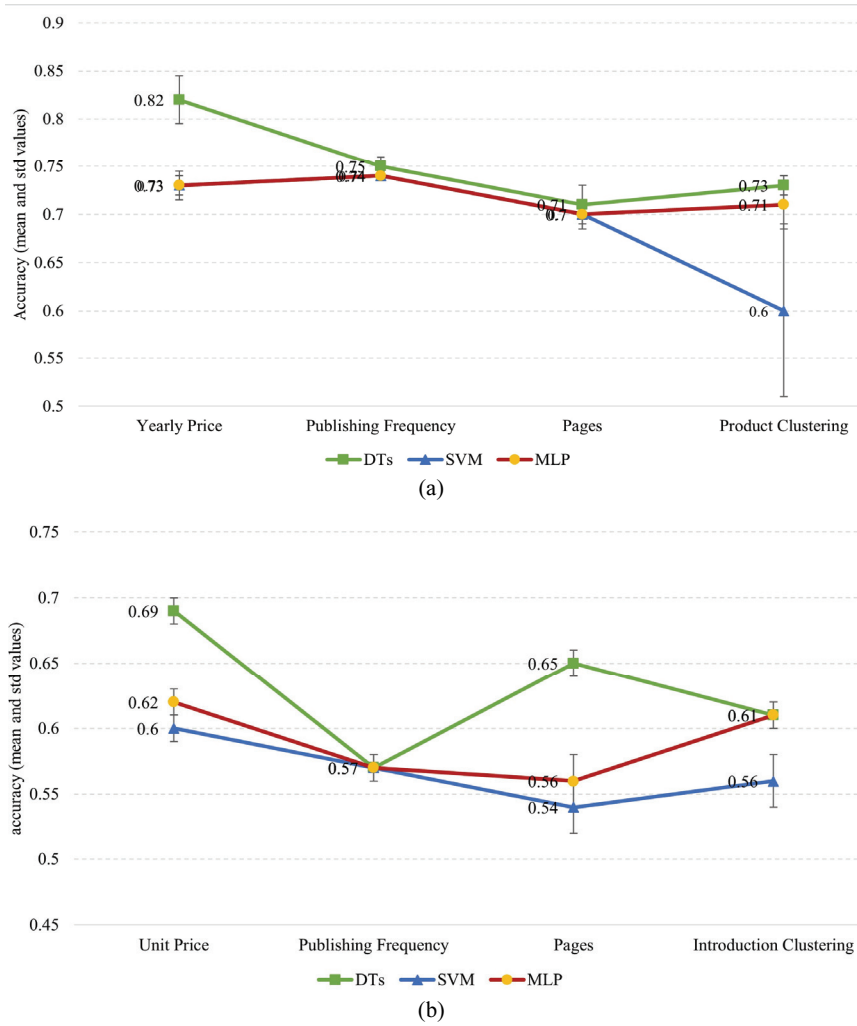

Fig. 4. Prediction accuracy in a single-factor case: (a) newspaper and (b) magazine.

4.2.1 Single-factor case

For newspapers and magazines, the predicted accuracies (mean and standard deviation) of each factor related to the best-selling status are presented in Fig. 4. Yearly price and unit price show better classification performance in newspapers and magazines, respectively.

4.2.2 Two-factor case

In the combination of two factors, as shown in Fig. 5, the accuracies of the price mixture (yearly prices or unit prices) or clustering-based feature mixture (product or introduction) were improved compared with those of the previous single-factor case. The combination of publishing frequency and number of pages also generally increased relative to the case of a single factor.

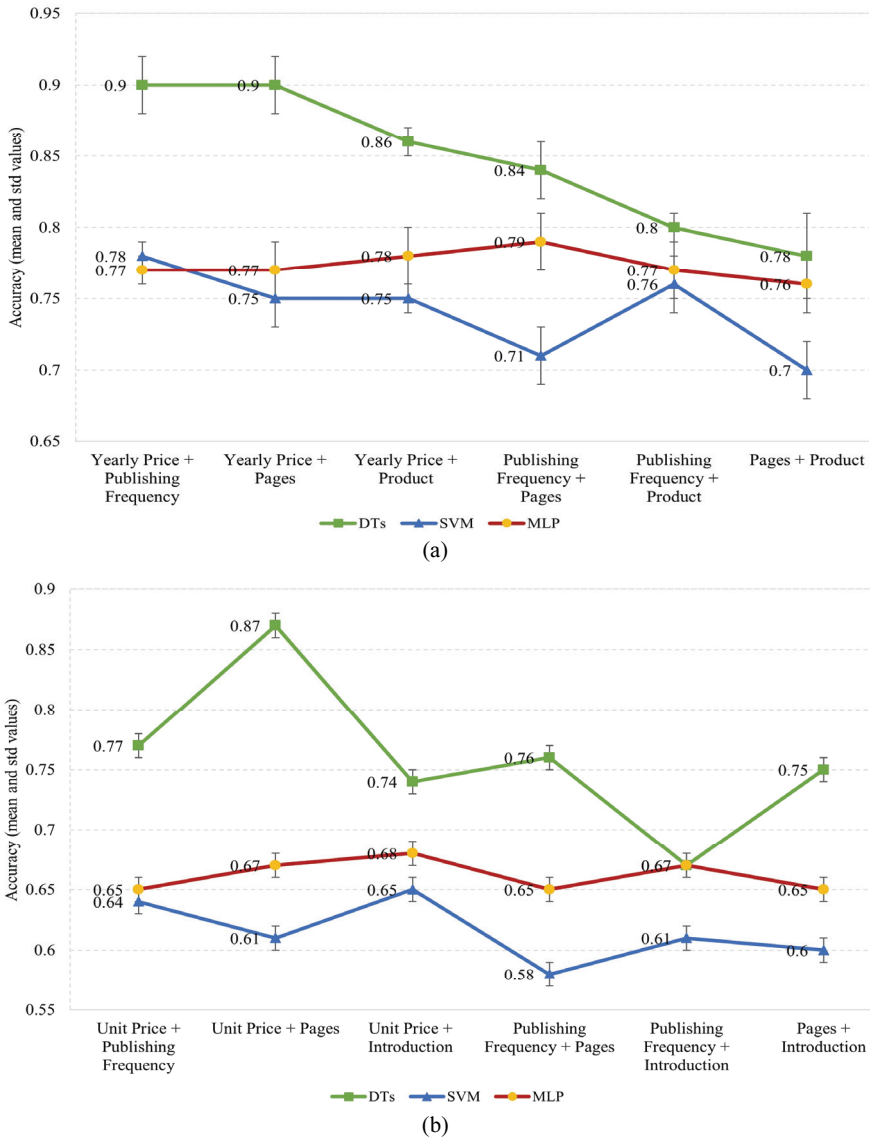

Fig. 5. Prediction accuracy in the case of two factors: (a) newspaper and (b) magazine.

4.2.3 Three-factor case

Fig. 6 shows that using a combination of three factors was more accurate than using a variety of two factors or a single factor.

4.2.4 Four-factor and all-factors cases

Four and all factors were considered as features for training and testing, and the accuracies obtained are illustrated in Fig. 7. The figure shows that using a combination of the four correlated factors yielded significantly better results than those of fewer combinations. When all elements were combined, the results slightly decreased for DTs and increased for SVM and MLP, especially in the magazine test.

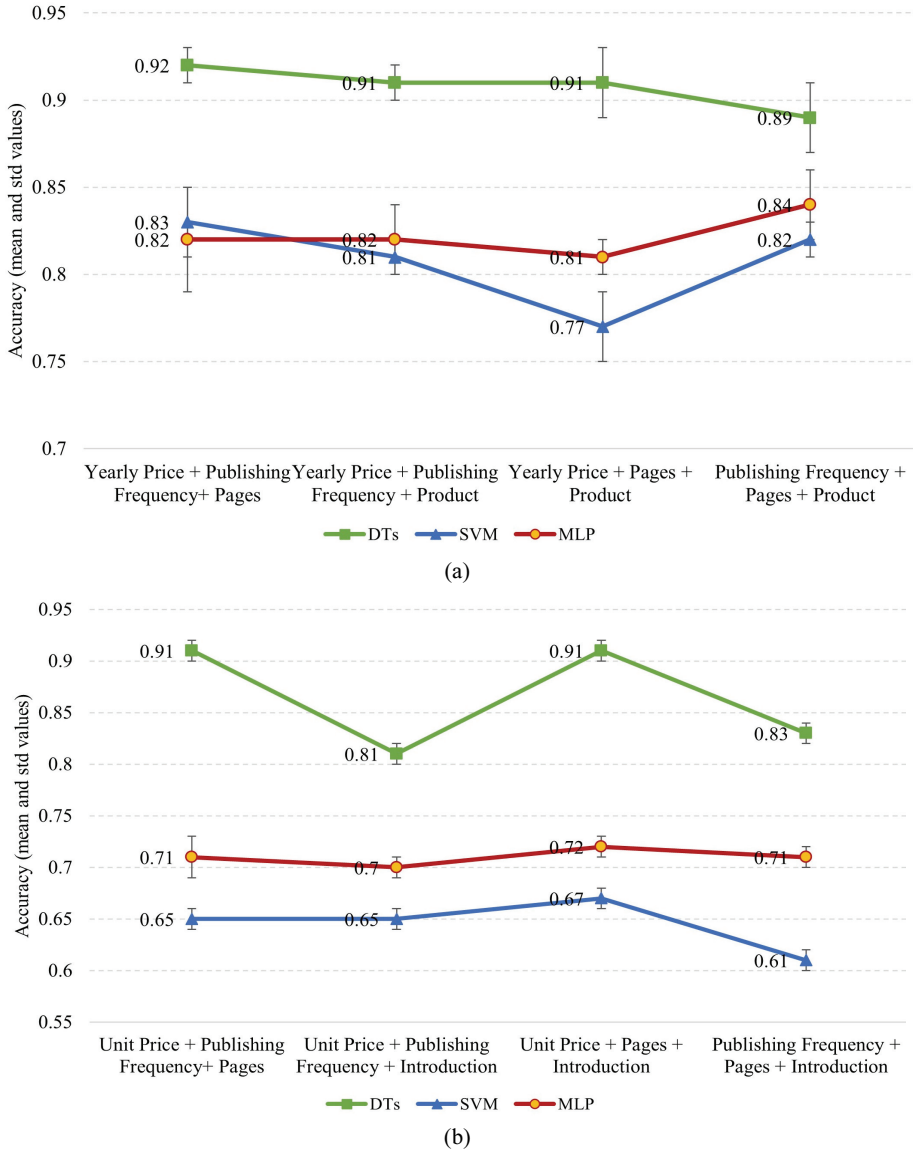

Fig. 6. Prediction accuracy in the case of three factors: (a) newspaper and (b) magazine.

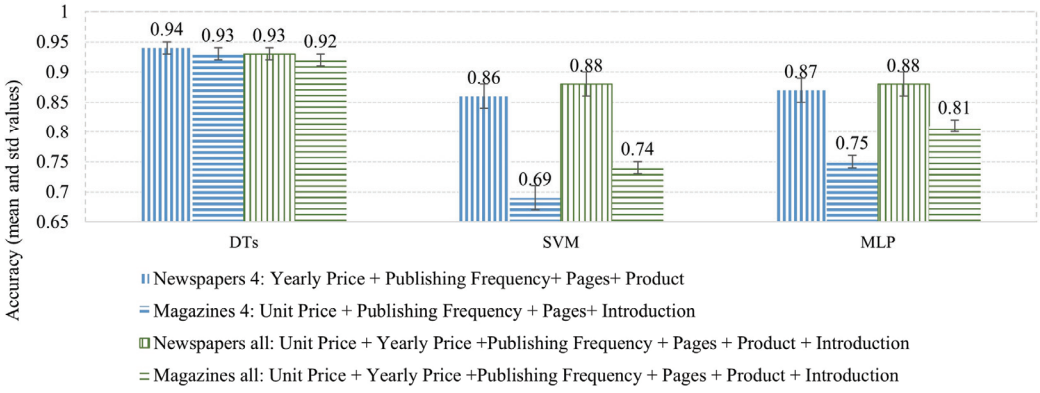

Fig. 7. Prediction accuracy in the case of four and all factors.

4.2.5 Discussion

As shown in Figs. 4–7, the corresponding prediction accuracy increases with the number of influencing factors. Under the same conditions, the accuracy of using DTs is considerably higher than that of using SVM and MLP. As the data dimensions increase, the performance of SVM and MLP improves. DTs are prone to overfitting; hence, when all the factors are used, the accuracy of using DTs decreases slightly, but that of SVM and MLP considerably improves. Using a random forest model can solve the problem of overfitting when all factors are considered. In all the combinations, the price factor shows a high importance rate. Cluster-based factors are not the worst performers and can improve the corresponding accuracy when combined with other factors; however, the single-factor performance is not outstanding. The main reason for the poor performance is that the data are too sparse during clustering.

Overall, we should focus on different combinations of factors for newspapers and magazines. For newspapers, yearly price and product category should be focused on, whereas unit price and introduction should be focused on magazines. Nearly 15.5% of the missing newspaper introduction values are replaced by the newspaper's name, resulting in an insignificant impact on newspaper sales. Meanwhile, an extremely broad product categorization of magazines results in a less significant effect on magazine sales.

5. Competitive Strategy

With the continuous deepening of industrialization, enterprise conglomeration, and marketization of news media, the newspaper industry must continuously improve its core competitiveness. In the new era of a rapid expansion of online media at minimal cost, the newspaper industry must explore and research connotations to pursue the coordination of “quality” and “quantity.”

- In terms of content, it is necessary to emphasize social benefits and implement newspapers and magazines to be close to reality, life, and the masses, and to improve the system and mechanism that integrate social and economic benefits.
- In terms of price, the principle of healthy competition should be followed; malicious low-price competition and predatory pricing should be avoided, although price discrimination can be used while different prices are charged to varying audiences in other regions.

- In terms of quality, the construction of high-quality content publishing and dissemination capabilities should be strengthened, and content carriers, methods, formats, systems, and mechanisms should be innovated.
- In terms of positioning, the content, style, and value orientation of the newspaper industry must be audience-oriented.

6. Conclusion

This paper reports on fundamental, statistical descriptive, and predictive analyses conducted by mining the postal catalogs of newspapers and magazines. Word frequency and cluster analyses were performed on profiles and product categories, combined with correlation analysis for constructing and selecting features. Finally, three classification models were trained and used to predict factors that affect the best-selling status of newspapers and magazines. As the dataset was small, the knowledge gained through data mining was limited. Moreover, similar relevant research and corresponding experimental control were lacking. Nevertheless, the data analysis scheme used in the study was feasible and provided complementary strategies for pricing, distribution, and content for newspapers and magazines to enhance their competitiveness.

References

[1] D. Y. Jiang, "Data Analysis on Postal Catalogue of newspapers and magazines," 2021 [Online]. Available: https://github.com/jinagdayou127333/Data_Analysis_News_magazines_2021.

[2] X. Liu and X Ma, "Primary study on the problems and skills in using the present directory for ordering journals by post," *Library Work in Colleges and Universities*, vol. 20, no. 3, pp. 58-61, 2000.

[3] C. Hu, A. Gong, and B. Zhu, "Analysis of national postal newspapers and periodicals in 2000," *Documentation, Information, & Knowledge*, vol. 2000, no. 2, pp. 73-74, 2000.

[4] Z. Xiang, "Research on newspaper pricing strategy," *Price: Theory and Practice*, vol. 2006, no. 7, pp. 36-37, 2006.

[5] Y. Xu, J. Cui, H. Xu, H. Lu, and J. Sheng, "Statistical analysis on the pricing of 619 university journals," *Chinese Journal of Scientific and Technical Periodicals*, vol. 25, no. 3, pp. 432-434, 2014.

[6] L. Peng, "Pricing strategy and analysis of professional scientific journals," *Chinese Journal of Scientific and Technical Periodicals*, vol. 27, no. 11, pp. 1121-1126, 2016.

[7] H. Bai, Y. Huang, L. Zhu, and Y. Zhu, "Analysis on the current situation and challenges of print media transformation," in *Proceedings of 2021 4th International Conference on Humanities Education and Social Sciences (ICHESS)*, Xishuangbanna, China, 2021, pp. 1650-1656. https://doi.org/10.2991/assehr.k.211220.279

[8] F. Wu, J. Qian, and K. Y. Liu, "Bottom line thinking and strategy reconstruction of newspaper distribution," *China Newspaper Industry*, vol. 2013, no. 17, pp. 32-33, 2013.

[9] H. Zhou, "40 Years of reform and development of China's press industry," *China Publishing Journal*, vol. 2018, no. 23, pp. 12-17, 2018.

[10] National Press and Publication Administration, "2019 News and Publication Industry Analysis Report," 2020 [Online]. Available: https://www.chinaxwcb.com/uploads/1/file/public/202011/20201104095548_wwrm2mol4a.pdf.

[11] X. Hu, "Research on the current situation and future trends of China's newspaper industry," *China Newspaper Industry,* vol. 2019, no. 15, pp. 34-39, 2019.

[12] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean Journal of Anesthesiology*, vol. 70, no. 4, pp. 407-411, 2017. https://doi.org/10.4097/kjae.2017.70.4.407

[13] Q. Zhang, X. Liu, and J. Fu, "Neural networks incorporating dictionaries for Chinese word segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 5682-5689, 2018. https://doi.org/10.1609/aaai.v32i1.11959

[14] X. Jin and J. Han, "K-means clustering," in *Encyclopedia of Machine Learning*. Boston, MA: Springer, 2011, pp. 563-564. https://doi.org/10.1007/978-0-387-30164-8_425

[15] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221-3245, 2014.

[16] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Heidelberg, Germany: Springer, 2009, pp. 1-4. https://doi.org/10.1007/978-3-642-00296-0_5

[17] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863-905, 2018. https://doi.org/10.1613/jair.1.11192

[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Support vector machines," in *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. New York, NY: Cambridge University Press, 2007, pp. 883-898.

[19] O. Tiryaki and C. O. Sakar, "Nonlinear feature extraction using multilayer perceptron based alternating regression for classification and multiple-output regression problems," in *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA)*, 2018, pp. 107-117. https://doi.org/10.5220/0006848901070117

[20] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics*, vol. 18, no. 6, pp. 275-285, 2004. https://doi.org/10.1002/cem.873

Dayou Jiang **https://orcid.org/0000-0001-5054-6958**

He is now a full-time teacher in the Department of Computer Science, Anhui University of Finance and Economics, China. He received a Ph.D. degree in copyright protection from Sangmyung University in 2020. He also received B.S. and M.S. degrees in Engineering College from YanBian University, China, in 2012 and 2016, respectively. His current research interests include multimedia processing, machine learning, and data science.