Open data policy, connectivity and portfolio determine brand popularity in e-commerce channel

Martijn J. Hoogeveen[1,2]

1. Department Management Sciences & Technology, Open University, the Netherlands.

2. Icecat Research, the Netherlands

Corresponding author: Martijn J. Hoogeveen, martijn.hoogeveen@icecat.com

Abstract

Content marketing is increasingly important for brand performance, i.e., brand popularity and associated e-commerce sales. Content marketing factors that are seen as effective for brand popularity in the e-ecommerce channel are open data syndication policies, connectivity to e-commerce platforms, reviews, data completeness, and a brand's online product portfolio.

In this study brand popularity, reviews and catalog data are used in combination with product reviews from a global panel of 3,057 e-commerce platforms (n = 333 manufacturing brands). Through stepwise backward multiple linear regression five highly significant predictive factors for online brand popularity are selected in our content marketing model: the brand's data syndication policy, the number of connected e-commerce platforms, a brand's number of products, its number of products per category and the number of product categories in which it is active. The model explains 78% of the variance of brand popularity and has a good and highly significant fit.

Keywords

Brand popularity, content marketing, predictive model, open data policy, e-commerce

Highlights

- Online brand popularity is critical for the success of manufacturers in the e-commerce channel.

- Content marketing predictors, jointly explain 78% of the variance in brand popularity.

- Online brand popularity is improved when manufacturing brands adopt an open content syndication policy, or have more connections to e-commerce platforms, more products per category or presence in more product categories.

- The added value of a brand's product review score or sustainability index as reliable predictors in our model is not yet clear.

Introduction

Content marketing (Rowley, 2008) is becoming increasingly important for online branding (Bamm et al., 2018) and influencing e-commerce sales (Geng et al, 2020). There are multiple aspects of content marketing that are seen as important. Online reviews on product usage experiences can be an effective way to build a strong brand image, which in turn helps to reduce consumer's buying uncertainty (Chakraborty & Bhat, 2018). Such user-generated content not only contributes to brand popularity, but also to a brand's sales (Lasmi et al, 2021). Further, product data completeness or information quality is believed to influence consumer purchase decisions as well (Amanah et al., 2018), particularly in e-commerce stores (Kumar & Ayodeji, 2021). Product data completeness can extend to adding congruent multimedia assets (Hoogeveen, 1997). Additionally, the size, breadth and depth of a brand's

portfolio is observed to have positive effects on sales (Kirca et al., 2020). And, finally content syndication has become a prominent element of content marketing in, for example, the publishing domain (Edo et al., 2019). Typical content syndication policies are "open" (Bruzzone et al., 2020) or "exclusive" (Chang & Jhang, 2020). These aspects can be understood in the light of the e-marketing mix (Sriram et al, 2019), the intersection of product, promotion, price and place related policies in the digital world.

The novelty of this research is in the aim to quantitatively assess the strength and significance of modern content marketing factors in relation to a manufacturer brands e-commerce performance, in particular brand popularity. Additionally, it is assessed to what degree these factors can satisfactorily explain the variance in brand popularity, or whether still other factors are needed. The ability to predict a brand's popularity based on content marketing variables, as a significant part of the e-marketing mix, is a powerful and modern tool for brand owners who want to quickly understand their e-commerce potential.

This study's hypothesis is that content marketing factors are significant predictors of a brand's popularity in the e-commerce sales channel. A further hypothesis is that a content marketing model, combining multiple significant factors, better explains the brand popularity of successful online manufacturing brands than each factor alone. Therefore, the main objective of this study is to assess content marketing factors, and the combination of factors, that best determine a brand's popularity in the e-commerce sales channel.

Methods

For this analysis, datasets are selected for an initial sample of 500 manufacturing brands, which is further reduced to 333 brands (n = 333) by taking into account only the overlapping brands from each dataset.

For e-commerce brand popularity, a brand popularity dataset (2022) from June 10, 2022, till September 24, 2022 is used, and the top 500 most successful brands are selected from this dataset. This dataset is collected based on a continuous and automated panel of 3,057 e-commerce platforms servicing retailers from 171 countries or regions – as defined by top level domains. The use of product information for 32,693 different manufacturing brands is real-time monitored. Over the given period of 90 days, per brand the following data are collected: product downloads, the number of connected e-commerce platforms, the content syndication policy of a brand, the number of products of a brand as an indication of portfolio size, and the number of categories in which a brand has at least 1 product as an indication of portfolio breadth. Product downloads is the number of times that a brand's products are downloaded by its e-commerce platforms or end-users in the given period of 90 days. The number of e-commerce platforms are the number of different e-commerce users that connect to the database to download product data-sheets of a brand in the defined sampling period. The syndication policy can be 'open', all of a brand's product data are made available as open data, or 'closed', the brand's product data is only available for selected or authorized e-commerce users.

Additionally, per brand a popularity rank is calculated on the basis of the product downloads metric, and the average number of products per category per brand are calculated as an indication of portfolio depth. A popularity rank of 1 is the best score, and a popularity rank of 500 is the lowest score in the dataset. As the open catalog is used by *10,000s of e-commerce companies from almost every country in the world on a daily basis for updating their own catalogs, it is seen as a sufficiently representative statistical sample. Finally,a a brand's data health score and brand review score are calculated. Data

health is an average based on the completeness of the product data-sheets which describe a brand's online products. The brand review score is calculated as the average of the review scores of all a brand's individual products and results in a figure between 0 and 100% (or 0 and 1). Of the selected 500 brands, 333 brands have product reviews in the database to calculate the average score from.

For sensitivity analyses, the brand popularity dataset is extended to include all 500 brands, initially selected. And, crude measures for brand popularity are compared with the brand rank transformation.

Variables are presented with their sample sizes (n), means (M), and standard deviations (SD). Correlation coefficients are calculated to assess the strength and direction of relations of each independent variable with brand popularity, and with each other.

Stepwise backward multiple linear regression for all independent variables on online brand popularity was used to keep only candidate predictors that are significant ($p < 0.05$) in the model and remove insignificant ones. During each step, per independent variable the (standard) coefficient, t-stat and its 95% confidence interval (CI), probability, and the variance inflation factor (VIF) value are calculated as well. Next, the predictors that were multicollinear are removed whereby we used an acceptable VIF value of at most 5 as a threshold, whereby a VIF value below 2.5 is seen as ideal. With the remaining independent variables the F-value, standard deviations and errors, degrees of freedom (DF), and significance level, are calculated to test the goodness of fit hypothesis for the predictive model for brand popularity. Finally, the multiple R, Multiple R squared (R^2) and adjusted R^2 correlation coefficients are calculated to estimate the predictive power of the model. On the basis of the regression outcomes, the algebraic equation to predict brand popularity is given, and a dot plot is provided to visualize the fit of the model's predicted versus the observed brand popularity values.

Additionally, as linear regression assumes normality of residuals, the Shapiro-Wilk test is applied, and to test the homoscedasticity requirement – homogeneity of the variance of residuals – the White test is

applied. Additionally, the priori power is calculated for each predictor separately and is compared with the outcomes of the predictive model.

For selected independent variables with a $p < 0.05$ and VIF score < 5, standard $\log_{10}$, and square root data transformations are applied to reduce non-linearity in relations between variables which helps to reduce skewness, and, especially, meet the normality and homoscedasticity requirements. Such data transformations do not change the nature and direction of relations between independent variables and brand popularity. As the variables in the datasets only contained positive numbers no further data transformations were necessary.

The results are reported in APA style.

All statistical analyses were done with Stats Kingdom 2022, which is benchmarked before on R version 3.5. Key outcomes are calculated in Excel as well as a check.

Results

The sample size (N), means, and SDs of the independent and dependent variables as used in the multiple linear regression models are given in Table 1. If applicable, the values are given for the data sets after applied data transformations.

Variable	N	Mean	SD
Downloads	333	9,237,924	3,130,1091
$Log_{10}(Platforms)$	333	1.83	0.46
Open data policy	333	0.46	0.50
$Log_{10}(Products/category)$	333	2.10	0.61
$Log_{10}(Categories)$	333	1.85	0.38
$Temperature^2$	333	221	142
Dew point temperature	333	8.56	5.70
Sqrt(Products)	333	130	155
Sqrt(brand popularity rank)	333	14.0	5.37

Table 1: Overview of mean (M), and standard deviation (SD) per independent variable as used in the multiple linear regression models. The function Sqrt returns the square root of the variable.

The independent variables that correlate (highly) significantly with online brand popularity (see Table 2), are in order of strength: platforms, open data policy, categories, products, and reviews score. The inverse nature of these correlations is logical as an improvement on these factors coincides with an improved, i.e., lower, brand popularity rank. For example, an open data syndication policy, more connections to e-commerce platforms, and a bigger catalog size (more products in more categories and more products per category) lead to a better, i.e., lower, brand popularity rank. The highly significant inverse correlation with downloads speaks for itself given that brand popularity rank is directly based on it. Unsurprisingly, the same factors correlate (highly) significantly with downloads, except for

reviews score. Worth to highlight is that data health correlates highly significantly with review score $(r(500) = 0.21$ $(p < 0.00001))$, especially given that these factors are fully independent measures.

	Brand popularity rank	Data Health	Downloads	Platforms	Review score	Open data policy	Products	Categories	PPC
Brand Popularity Rank	1.000	0.054	**<u>-0.351</u>**	**<u>-0.538</u>**	**-0.102**	**<u>-0.321</u>**	**<u>-0.261</u>**	**<u>-0.320</u>**	-0.047
Data Health	0.054	1.000	-0.029	-0.035	**<u>0.207</u>**	0.060	-0.002	-0.064	**-0.112**
Downloads	**<u>-0.351</u>**	-0.029	1.000	**<u>0.669</u>**	0.018	**<u>0.176</u>**	**<u>0.648</u>**	**<u>0.432</u>**	0.038
Platforms	**<u>-0.538</u>**	-0.035	**<u>0.669</u>**	1.000	0.075	**<u>0.522</u>**	**<u>0.313</u>**	**<u>0.438</u>**	-0.067
Reviews score	**-0.102**	**<u>0.207</u>**	0.018	0.075	1.000	0.045	0.033	**-0.090**	0.054
Open data policy	**<u>-0.321</u>**	-0.060	**<u>0.176</u>**	**<u>0.522</u>**	**<u>0.045</u>**	1.000	-0.012	0.070	-0.084
Products	**-0.261**	-0.002	**<u>0.648</u>**	**<u>0.313</u>**	0.033	-0.012	1.000	**<u>0.225</u>**	**<u>0.297</u>**
Categories	**<u>-0.320</u>**	-0.064	**<u>0.432</u>**	**<u>0.438</u>**	**-0.090**	0.070	**<u>0.225</u>**	1.000	**-0.104**
Products/Category	-0.047	**-0.112**	0.038	-0.067	0.054	-0.084	**<u>0.297</u>**	**-0.104**	1.000

Table 2: Overview of correlations between variables before transformations and limitations of the dataset (n = 500). Bold: significant at p < 0.05. Bold+Underlined: highly significant at p < 0.01.

After transformations and dataset restriction to brands with a review score, the correlations with brand popularity are more or less the same. Only the correlation with products per category has become highly significant as well $(r(333) = -0.46, p < 0.00001)$.

In both situations there are no collinears, given that no pair of independent variables have a strong correlation $(r \geq 0.8)$ with each other.

After multiple iterations during stepwise backward multiple linear regression, five independent variables were selected from the set of online marketing variables that are both significant $(p < 0.05)$ and have an acceptable VIF value below 5. These selected predictors are: open syndication policy, the

number of connected e-commerce platforms, the number of products, the products/category and the number of categories in which a brand has products (see Table 3). In the total datasets, Data Health did not have a significant correlation with brand popularity ($r(500) = 0.06$, $p = 0.23$) and was deselected for the model. Also, Product Reviews was deselected, despite a significant correlation with brand popularity ($r(500) = -0.10$, $p = 0.022$) and given that it did not add explanatory power to the model.

	Coeff.	SE	t-stat	lower $t_{0.025(327)}$	upper $t_{0.975(327)}$	Stand. Coeff.	P	VIF
b	51.7	1.63	31.6	48.5	54.9	0	<0.00001	
Log$_{10}$(Categories)	-7.26	0.55	-13.2	-8.34	-6.18	-0.51	<0.00001	2.20
Log$_{10}$(Platforms)	-4.64	0.42	-11.2	-5.45	-3.82	-0.40	<0.00001	1.91
Open data policy	-2.47	0.35	-7.14	-3.16	-1.79	-0.23	<0.00001	1.55
Sqrt(Products)	0.00727	0.00166	4.37	0.00399	0.0105	0.209	0.000017	3.44
Log$_{10}$(Products/Category)	-7.40	0.42	-17.5	-8.23	-6.57	-0.84	<0.00001	3.43

Table 3: Overview of outcomes per predictor after multiple linear regression without downloads resulting in an adjusted R-squared = 0.778. Selection of predictors is based on being (highly) significant and having multicollinearity (VIF) score below 5. The function Sqrt returns the square root of the variable.

On the basis of this test, the null-hypothesis (H_0) can be rejected that the predictive content marketing model with the five selected factors does not provide a good fit: $F_{(5, 327)} = 233.5$, $p < 0.00001$. R^2 equals 0.781, which means that the predictors explain 78.1% of the variance of brand popularity (Adjusted R square = 0.778 and R = 0.884) (see Fig. 1).
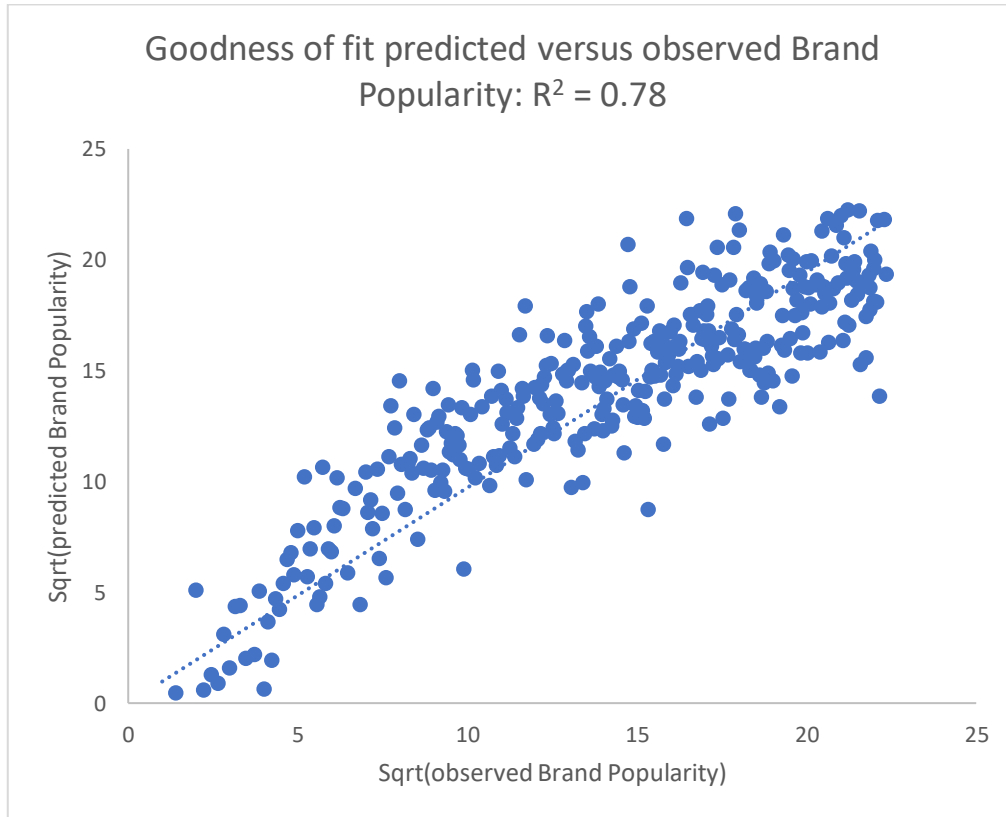
Fig. 1: The predictive content marketing model is explaining 78.1% of the variance of the observed Brand Popularity.

Below, the predictive content marketing model's regression formula for $\widehat{BPR}$:

$$\left(51.7 - 7.25 \log_{10} C - 4.64 \log_{10} Pl - 2.47 Po + 0.00727 \sqrt{Pr} - 7.40 \log_{10} PPC\right)^2$$

Where $\widehat{BPR}$ is the predicted brand popularity rank, C is the number of categories in which a brand has products, Pl is the number of e-commerce platforms that are downloading a brand's product data, Po indicates whether a brand has an open data ("1") syndication policy or an exclusive one ("0"), Pr is the number of products that a brand has in total in its catalog, and PPC is the number of products that a

brand has per category in its catalog. *C, Pr, and PPC* are all variables describing the size (breadth and depth) of a brand's product portfolio.

Given the somewhat raised multicollinearity scores, *Products* has to be interpreted in the content marketing model as a correction on *PPC*.

As a sensitivity analysis crude downloads are used as a dependent variable instead of brand popularity rank. It leads to similar though not completely satisfactory outcomes because of methodological concerns related to the homoscedasticity of residuals and normality requirements. When including the whole dataset (n = 500 brands) despite having quite some missings does lead to the inclusion of the brand review score in the predictive model but again homoscedasticity or normality requirements cannot be satisfactorily met.

Discussion

The predictive power of the content marketing model including a brands connected platforms, its content syndication policy, the number of products, products per category and the number of categories in which it is active is good (77.4%). For example, an open syndication policy, more connected e-commerce platforms, and more products per category and more categories in which a brand is active, are related to a better (lower) brand popularity score. The number of products can be seen as a correction on the product per category variable in the model.

When the model is tested by using the number of downloads of these products as *independent* variable, the predictive power of the model remains more or less the same. Given that the downloads factor was used to determine the brand popularity rank, it implies that the content marketing model contains a robust – equally powerful - set of independent predictors. The transformation of downloads into brand popularity rank is superior to other transformations in the sense that homoscedasticity and

normality requirements are fully met, which is not the case when downloads is used as *dependent* variable instead of brand popularity rank.

Assuming that the same online marketing factors are relevant as predictors for brands outside the sample, it could be hypothesized that the model has a similar predictive power for brands in other sectors or categories outside the sample, or, to a certain degree, for non-manufacturing brands.

Methodological concerns

The online brand popularity panel is relatively strong in Western countries, which might lead to a sampling bias. It would be interesting to see if the derived predictive content marketing model would fundamentally change if the panel is strengthened in East-Asian and African countries.

Further, it would of interest to see the completeness of the product reviews dataset in terms of brand coverage being improved. It might lead to the inclusion of product reviews score as a significant predictor, as shown in a sensitivity analysis, leading to a slightly improved predictive power of the model.

Finally, it would still be of interest to see if the predictive power of the model can be improved by adding other predictors. For example, in addition to portfolio (product) and connected platforms (place), also the aspects of average price and promotion as part of the e-marketing mix. Also of interest is whether a brand's sustainability score has added value for the predictive model. A brand's sustainability reporting (Loh & Tan, 2020) is increasingly seen as impactful regarding corporate performance (Cowan & Guzman, 2020), especially given the observation of a clear relation between e-commerce and sustainability (Reijnders & Hoogeveen, 2001).

Conclusion

A predictive content marketing model explains 78.1% of the variance of online brand popularity of a sample of 333 successful manufacturing brands, and has a good fit ($F_{(5, 327)} = 233.5$, $p < 0.00001$), as the predicted and observed brand popularity correlate strongly and highly significantly. The significant predictors in the content marketing model are a brand's content syndication policy, the number of connect e-commerce platforms, a brand's number of products, products per category and the number of categories in which it is active. The mentioned factors are inversely correlated with brand popularity rank, i.e., an open syndication policy, more connected e-commerce platforms, more products per category and being active in more categories all lead to a better (lower) brand popularity rank. In the model, the number of products, another attribute of a brand's catalog size, should be seen as a correction on the product per category variable.

Using brand popularity rank as a dependent variable leads to more statistically robust models than when using crude popularity values, i.e., product data-sheet downloads, as an online brand popularity indicator.

It is yet to be determined whether a brand's product review score, sustainability rating, or pricing policy has added value for a predictive content marketing model.

References

Amanah, D., Hurriyati, R., Gaffar, V., Layla, A. A., & Harahap, D. A. (2018, December). Effect of price and product completeness to consumer purchase decision at Tokopedia. com. In Proceedings of the 2nd Global Conference on Business, Management and Entrepreneurship 2017 (Vol. 1, pp. 34-37).

Bamm, R., Helbling, M. and Joukanen, K. (2018), "Online Branding and the B2B Context", Koporcic, N., Ivanova-Gongne, M., Nyström, A.-G. and Törnroos, J.-Å. (Ed.) Developing Insights on Branding

in the B2B Context, Emerald Publishing Limited, Bingley, pp. 163-176. https://doi.org/10.1108/978-1-78756-275-220181009.

Bruzzone, A. G., Agresta, M., & Hsu, J. H. (2020). Word of Mouth, Viral Marketing and Open Data: A Large-Scale Simulation for Predicting Opinion Diffusion on Ethical Food Consumption. International Journal of Food Engineering, 16(5-6).

Chakraborty, U. and Bhat, S. (2018), "Credibility of online reviews and its impact on brand image", Management Research Review, Vol. 41 No. 1, pp. 148-164. https://doi.org/10.1108/MRR-06-2017-0173.

Chiang, I. R., & Jhang-Li, J. H. (2020). Competition through exclusivity in digital content distribution. Production and Operations Management, 29(5), 1270-1286.

Cowan, K., & Guzman, F. (2020). How CSR reputation, sustainability signals, and country-of-origin sustainability reputation contribute to corporate brand performance: An exploratory study. Journal of business research, 117, 683-693.

Edo, C., Yunquera, J., & Bastos, H. (2019). Content syndication in news aggregators: towards devaluation of professional journalistic criteria= La sindicación de contenidos en los agregadores de noticias: Hacia la devaluación de los criterios profesionales periodísticos.

Geng, R., Wang, S., Chen, X., Song, D., & Yu, J. (2020). Content marketing in e-commerce platforms in the internet celebrity economy. Industrial Management & Data Systems, 120(3), 464-485.

Hoogeveen, M. (1997). Toward a theory of the effectiveness of multimedia systems. International journal of human-computer interaction, 9(2), 151-168.

[dataset] Icecat. Brand popularity rank search and overview page. https://icecat.biz/en/brands. Accessed September 25, **2022**. DOI: 10.13140/RG.2.2.19627.41763

Kim, M.-Y., Moon, S., & Iacobucci, D. (2019). "The Influence of Global Brand Distribution on Brand Popularity on Social Media". Journal of International Marketing, 27(4), 22–38. https://doi.org/10.1177/1069031X19863307.

Kirca, A. H., Randhawa, P., Talay, M. B., & Akdeniz, M. B. (2020). The interactive effects of product and brand portfolio strategies on brand performance: Longitudinal evidence from the US automotive industry. International Journal of Research in Marketing, 37(2), 421-439.

Kumar, V. & Ayodeji, O.G. E-retail factors for customer activation and retention: An empirical study from Indian e-commerce customers, Journal of Retailing and Consumer Services, Vol. 59, 2021. https://doi.org/10.1016/j.jretconser.2020.102399.

Lasmi, H., Lee, C. H., & Ceran, Y. (2021). Popularity Brings Better Sales or Vice Versa: Evidence from Instagram and OpenTable. Global Business Review, 09721509211044302.

Loh, L., & Tan, S. (2020). Impact of sustainability reporting on brand value: an examination of 100 leading brands in Singapore. Sustainability, 12(18), 7392.

Qian, L., Dou, Y., Xu, X., Ma, Y., Wang, S., & Yang, Y. (2022, August). Product Success Evaluation Model Based on Star Ratings, Reviews and Product Popularity. In 2022 8th International Conference on Big Data and Information Analytics (BigDIA) (pp. 295-302). IEEE.

Reijnders, L., & Hoogeveen, M. J. (2001). Energy effects associated with e-commerce: A case-study concerning online sales of personal computers in The Netherlands. Journal of Environmental Management, 62(3), 317-321.

Rowley, J. (2008). Understanding digital content marketing. Journal of marketing management, 24(5-6), 517-540.

Sriram, K. V., Phouzder, K., Mathew, A. O., & Hungund, S. (2019). Does e-marketing mix influence brand loyalty and popularity of e-commerce websites? ABAC Journal, 39(2).