World Journal of Economics and Business Research
ISSN: 2960-0081
DOI: https://doi.org/10.61784/wjebr3016

LEVERAGING PRICE AND TIME DATA TO ENHANCE USER RATING PREDICTIONS OF AMAZON BOOK: BASED ON KAGGLE DATA

Bidhan Dolai[1*], Indrajit Ghosh[2]
[1]*Research Scholar, SGBAU, MH, India.*
[2]*Libary Attendant, Baba Saheb Ambedkar Education University, West Bengal, India.*
Corresponding Author: Bidhan Dolai, Email: bidhandolai93@gmail.com

Abstract: Drawing on the existing literature pertaining to the online marketplace, it can be seen that in such a marketplace, consumer ratings of products are the most important predictors of their success and customer satisfaction. This paper, therefore, drew on data sourced from Kaggle and focusing on Amazon books from the year 2009 to 2019, examined the correlations among three main variables, namely Year of release, Price, and the number of Reviews and User Ratings. It was established through multiple regression analysis that Price and Year markedly affect User Ratings while Reviews had a marginally significant effect. An explanation could be that with an increase in price, the users tend to give low ratings which suggests that there is a likelihood of dissatisfaction amongst the consumers, as opposed to older books which generally tend to score lower user ratings because of existing demand for newer works. Considering the R^2 value of 0.074, only 7.4% of the variation in User Ratings can be accounted for by the given variables, thus, more predictors will be required in subsequent investigations. Essentially, the present study sheds light on factors affecting user's satisfaction, in particular, the role of pricing and the timing of releases in satisfying the users.
Keywords: Kaggle data; Rating prediction; Data analytics

1 INTRODUCTION

In the swiftly evolving virtual marketplace, consumer scores have come to be an essential indicator of product success and client pride across various industries. Online structures, e-trade web sites, and digital content vendors increasingly depend on those ratings to inform purchaser decisions, enhance product visibility, and manual strategic business initiatives. Understanding the determinants of user rankings is important for businesses aiming to optimize their services and align with consumer preferences. This observe investigates the connection between 3 key variables—Year (representing the e-book or release date), Price, and Reviews—and their impact on User Ratings thru a a couple of regression analysis [1].

Previous studies have highlighted the large roles that pricing techniques, the timing of product releases, and the volume of purchaser critiques play in shaping consumer pride and ratings. Price can have an effect on consumer perceptions of affordability and competitiveness, potentially affecting their basic pleasure with a product or service [2]. Year, as a trademark of the product's release date, may mirror its relevance, technological advancements, and alignment with modern market traits, thereby impacting consumer scores. Reviews, encompassing the quantity and nature of purchaser feedback, serve as a shape of social proof and can appreciably sway potential customers' perceptions and selections [3].

Despite the diagnosed significance of those elements, the extent to which Year, Price, and Reviews collectively provide an explanation for versions in User Ratings remains insufficiently explored. This has a look at employs a regression model to quantify the contributions of these variables, aiming to elucidate their character and combined effects on person ratings. By studying these relationships, the research seeks to identify which factors are most influential in figuring out consumer delight and the way they have interaction to shape typical rankings [4].

The regression analysis reveals that Price and Year are statistically considerable predictors of User Ratings, whilst Reviews method importance. Specifically, higher costs are related to moderate decreases in person scores, suggesting that clients can also perceive more highly-priced merchandise as less satisfactory if their expectancies are not met. Conversely, newer merchandise tends to get hold of higher person rankings, indicating that current releases may additionally benefit from greater capabilities, stepped forward high-quality, or extra alignment with current consumer needs. The marginal significance of Reviews suggests that while customer remarks volume performs a role in influencing scores, its impact can be less reported in comparison to Price and Year [4].

However, the version's explanatory strength, as indicated by way of an R Square fee of 0.074, shows that handiest 7. 4 % of the variability in User Ratings is accounted for via Year, Price, and Reviews. This low R Square price highlights the presence of different influential elements no longer captured within the modern model, consisting of product first-rate, emblem recognition, advertising efforts, or client demographics. Consequently, even as Year and Price provide treasured insights into consumer ratings, there is a clear want for incorporating additional predictors to decorate the model's robustness and comprehensiveness [5].

In end, this study contributes to the information of how Year, Price, and Reviews have an impact on User Ratings, supplying a basis for destiny research to discover a more considerable set of variables. By identifying the sizable factors that affect person satisfaction, companies can better tailor their techniques to satisfy customer expectancies, optimize pricing systems, and effectively interact with their customer base. Ultimately, a deeper understanding of those dynamics is important for fostering high quality user reviews and reaching sustained market fulfillment.

2 LITERATURE REVIEW

Traditional CF approaches rely much on numerical evaluation to measure the similarity of items. Therefore, they tend to ignore the linguistic nuances of user preferences [6]. New findings indicate that converting absolute ratings into probabilistic linguistic terms could be used in order to improve the depiction of user sentiment and feature importance The inclusion of Bhattacharya coefficient enables semantic similarity to be tuned based on how the users are behaving. This way, the prediction becomes much more reliable.

The prediction of automobile scores has increasingly become more important for the optimization of vehicle layout and improving buyer attraction [7]. Traditional techniques usually rely on single-modal data, in the form of words or pictures, leading to a narrow analysis that misses the vast scope of information that exists [7]. Such a problem may lead to inaccuracies or avoid improvements within the car sector. Multimodal systems research frameworks integrate statistics from multiple sources, including parametric definitions, text-based descriptions and images, which remarkably improves the accuracy of prediction. Researchers found that multimodal models can achieve increments in explanatory power of 4% to twelve percent relative to unimodal peers [7]. In addition, through utilizing equipment such as SHAP for sensitivity analysis, interpretability is improved with actionable insights for designers and makers [7]. This shift towards a multi-modal approach is central to the progress in car design, review, and average market performance.

Techniques such as user-based CF have been widely applied in recommendation systems to predict ratings of unrated items from the resultant user interactions. Conventional similarity measures, however, tend to disregard the contributions of users and are usually limited by the coverage rating predictions, according to Kim in 2023. A new similarity metric, therefore, has been developed that integrates the degree of user contribution to predictions, which helps to fill that gap. It contributes towards better coverage and strengthens the recommendations overall. The method further uses item weighting in respect to rating frequency, thus increasing the scope of predictable items. Extensive experiments performed on benchmark datasets showed that this new measure improves the quality of recommendations, increases diversity, and hence contributes to a more robust recommendation system.

3 OBJECTIVES

(1)Examine how the price affects Amazon book reviews from users.
(2)Look at the connection between user ratings and the year of release.
(3)Evaluate how many reviews have an impact on user ratings.

4 METHODOLOGY

The methodology for this study involved analyzing a dataset sourced from Kaggle, which included Amazon book listings from 2009 to 2019. The key variables included User Ratings (dependent variable), Year of release, Price, and the total number of Reviews (independent variables). Data cleaning was conducted to address any missing or inconsistent entries, ensuring the integrity of the dataset. A multiple regression analysis was then performed to quantify the relationships between the independent variables and User Ratings. The model's performance was evaluated using R2, adjusted R2, and ANOVA to assess the statistical significance of the predictors. The coefficients obtained from the regression analysis were interpreted to elucidate the nature and strength of the relationships, ultimately providing insights into how Year, Price, and Reviews influence User Ratings in the digital marketplace.

5 DATA ANALYSIS AND INTERPRETATION

Table1 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.272[a]	.074	.069	.2190

a. Predictors: (Constant), Year, Price, Reviews

A regression analysis was performed to investigate the impact of Year, Price, and Reviews on User Rating (Table 1). The overall model established was productive in terms of predicting the user rating, F (3,546) =14.501, p<.001 F (3,546) =14.501, p<.001derives that the predictors in general significantly explain the differences observed in User Ratings across

the demographic. The correlation coefficient *R* R was established at 0.272 this indicates a weak relationship between the predictors combined and the User Rating. Correspondingly, the obtained R Square value of 0.074 indicates only 7.4% of User Ratings is explained by the Year, Price, and Reviews, while the geered R Square of 0.069 also confirms the low explanations due to the number of predictors in the model [8]. The standard error in the regression was found to be 0.2190, which shows the extent to which observed User Ratings deviate from their predicted values. Further, although the model appears to be acceptable, it is convincingly observable that it is grossly deficient in R Square more so with factors which were not included in the analysis are and will be the most potent variables. These results point out the necessity to add more diverse predictors into the model for better statistical results in future studies [9].

Table 2 ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.087	3	.696	14.501	.000[b]
	Residual	26.197	546	.048		
	Total	28.285	549			

a. Dependent Variable: UserRating
b. Predictors: (Constant), Year, Price, Reviews

Analysis of variance (ANOVA) for regression models assessing the impact of year, price, and reviews on user ratings. revealed a significant overall effect (Table 2). F (3,546) =14.501, p<.001F (3, 546) = 14.501, p < .001F (3,546) = 14.501, p < .001 The sum of squares regression (SSR) is 2.087, which indicates the amount of Variance in user ratings explained by the predictor. On the contrary the residual sum of squares (SSE) was 26.197, which represents the variance not accounted for in the model. The sum of squares (SST) is 28.285, which is the total variation of user ratings found in the dataset. A significant F statistic indicates that this model provides a better fit to the data than a model that does not. Predictor This confirms that at least one of the predictors (year, price, or review) significantly contributes to explaining the variation in user ratings. However, considering the relatively low R Square value (0.074) from the model summary, It is clear that although this model is statistically significant, But it explains only a small proportion of the total variation in user ratings [10].

Table 3 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-31.013	6.219		-4.987	.000	-43.229	-18.797
	Reviews	-1.512E-6	.000	-.078	-1.826	.068	.000	.000
	Price	-.002	.001	-.104	-2.478	.013	-.004	.000
	Year	.018	.003	.247	5.736	.000	.012	.024

a. Dependent Variable: UserRating

The focus of the regression analysis in this particular study was on Reviews, Price, and Year, each examined separately for its contribution to User Rating (Table 3). The users did not rate the Price, t(92) = -2.478, p = .013β = -0.104, This, too, was observed to be significant where Prices charged (β =−0.104,t =−2.478,p =.013,¥beta = -0.104, t = -2.478, p = .013 these situations of] causing a loss) formed the highest relationship. Also, Year β= 0.247, t = 5.736 p<.001β = 0.247, It was shown that, what this means is that with every unit increase in Price, the value of the User Rating decreases by 0.002 units holding also other variables constant. Such an approach is pertinent as variable Year exhibit a strengthening correlation, as every new year pursues an increase of the User Rating by 0.018 units holding the Price and the Reviews constant. The Reviews variable approach β=−0.078, t=−1.826, p=.068¥beta = -0.078 t=-1.826 p=.068 0.68 this compares the net effect of making people to write. But the review bias effect was not found, turning to confidence intervals holds however evidence that users at t=1.826 p<.068, And furthermore the two horizons enhanced the modelling p = 0.00552. The constant of -31.013 is significant (B = – 31.013, p < .001 Number n in Figure. 2 approaches and pursued advantages or other ends where technicians manipulated relation level or desire level.

6 RESULTS

A regression analysis was performed to determine their influence on User Ratings of Year, Price, and Reviews. The model had a statistically significant overall effect, F (3,546) =14.501, p <. 001F (3,546) = 14.501, p <. 001F (3,546) =14.501, p<. 001); Thus, the predictors together explain part of the variability in user ratings but the correlation coefficient (R=0.272), reveals that combined predictors have a weak connection to ratings from users. The low R2 (= 0.074) suggests that the model explains just 7.4% of the variance in user ratings, which indicates likely sources of additional rating influences not present within the analysis. Price was significant predictor with a β=−0.104, t=−2.478, p=. 013 beta = -0.104, t = -2.478, p = 013β=−0.104, t=−2.478, p=. 013. It implies that, with the other variables being held constant, for every increase in price of 1 unit there would be a decrease in user ratings by 0.002 units. Year was also significant and positive, (β=0.247, t=5.736, p<. 001 beta = 0.247, t = 5.736, p <.001 001β=0.247, t=5.736, p<. 001, indicating that the total difference in user ratings can be more than 0.018 unit between each 1 year of release by lack of new period due to few products with high age tend to have low ratings yet, meaning newer products are generally rated higher.

7 CONCLUSION

In summary, the analysis above suggests that Price and Year are each a positive predictor of User Ratings, while Reviews have a smaller effect on it. While the regression model is statistically significant, an R2 squared of 0.074 shows that it explains only a small piece of variation in user reviews suggesting that there are other factors at play. Findings suggest the value of autodetection and other predictors as a wider range of features is necessary in future studies to improve model interpretability. The Price showed a significant negative relationship with user rates and, therefore, should consider careful pricing strategies to avoid alienation of the users with their satisfaction of the software products.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE:

[1] Hong C, & Zhu K. ENHANCING YELP RATING PREDICTIONS. 2021. https://charleshong3.github.io/projects/cs182_yelp.pdf.
[2] Tayfur G. Real-time flood hydrograph predictions using rating curve and soft computing methods (GA, ANN). Handbook of Hydroinformatics, 2023.
[3] Angelini G, Candila V, Angelis L D. Weighted Elo rating for tennis match predictions. European Journal of Operational Research, 2022, 297(1): 120–132. DOI: https://doi.org/10.1016/j.ejor.2021.04.011.
[4] Kutlimuratov A, Abdusalomov A, Whangbo T K. Evolving hierarchical and tag information via the deeply enhanced weighted non-negative matrix factorization of rating predictions. Symmetry, 2020, 12(11). https://www.mdpi.com/2073-8994/12/11/1930.
[5] Lin Y, Ren P, Chen Z, et al. Meta matrix factorization for federated rating predictions. 2020. DOI: https://doi.org/10.1145/3397271.3401081.
[6] Jiangzhou D, Songli W, Qi W. A linguistically asymmetric similarity decision model integrating item tendency for rating predictions, 2024. DOI: https://doi.org/10.1177/01655515231220172.
[7] Saaidin S, Kassim M. Recommender system: Rating predictions of steam games based on genre and topic modelling. 2020 IEEE International Conference, 2020. https://ieeexplore.ieee.org/abstract/document/9140194/.
[8] Giovanni A, Candila V. Weighted Elo rating for tennis match predictions. European Journal of Operational Research, 2021, 297(1), 120–132. https://iris.uniroma1.it/handle/11573/1556145.
[9] Zuva T, Zuva K. Grouping recommender system users in distinct technology diffusion classifications for rating predictions. Academia, 2020. https://www.academia.edu/download/62182979/AC-6420200224-111807-r5uqw9.pdf.
[10] Ashokan A, Haas C. Fairness metrics and bias mitigation strategies for rating predictions. Information Processing & Management, 2021. https://www.sciencedirect.com/science/article/pii/S0306457321001369.