Gyan Manthan
Journal homepage: www.gyanmanthan.org ISSN: **3108-3455**

Optimizing Publication Outcomes Using Predictive Modeling: A Data-Driven Approach to Classifying High-Impact and Low-Impact Journal Papers in Social sciences, Business, Healthcare and AI

Hardi Amin[1], Zainab Siddiqi[2]

Abstract

Academic publishing plays a crucial role in knowledge dissemination, academic recognition, institutional reputation, and career advancement. Selecting an appropriate journal for publication is a major challenge for researchers because incorrect journal targeting often results in rejection, publication delays, and reduced research visibility. The present study examines whether measurable structural characteristics of academic papers can classify them as belonging to high-impact or low-impact journals before submission. Using predictive analytics and machine learning techniques, the study analyzed 104 academic papers collected from databases such as ScienceDirect, IEEE, Emerald, JMIR, Taylor & Francis, and MDPI across domains including Business, Healthcare, Social Sciences, and Artificial Intelligence. Variables such as page count, word count, citation intensity, author count, visual density, paper type, and number of references were considered as predictors, while journal impact factor served as the target variable. Four classification models—Generalized Linear Model (GLM), Logistic Regression, Decision Tree, and Random Forest—were evaluated using accuracy, specificity, AUC, confusion matrix, and lift chart analysis. The findings revealed that Logistic Regression achieved the best overall performance with high specificity, indicating strong capability in identifying low-impact papers. Citation intensity, visual density, paper length, and AI-related domains emerged as significant predictors of publication impact. However, the models demonstrated limited sensitivity in identifying true high-impact papers, suggesting that qualitative factors such as novelty, writing quality, and author reputation also influence publication success. The study concludes that predictive analytics can serve as a valuable decision-support tool for researchers and institutions in improving journal targeting strategies and optimizing publication outcomes.

Keywords: Academic Publishing, Predictive Analytics, Machine Learning, Journal Impact Factor, Logistic Regression, Citation Intensity, Visual Density, Research Papers, Scholarly Communication, Journal Targeting, Publication Strategy, High-Impact Journals, Bibliometrics, Research Visibility.

Introduction

Academic publishing is one of the most important practices for sharing knowledge, institutional prestige and professional recognition for academicians and researchers. Publications specifically in high-impact journals can influence academic promotion, funding

opportunities, scholarly visibility and university rankings **Error! Reference source not found.**. Academic publishing is a billion-dollar industry with lots of competition for placement in high impact journals.

Misaligned submissions lead to long delays (3-6month review cycles), wasted effort revising and resubmitting, reduce research visibility and possible increased administrative burden. With all of this in mind, choosing the proper journal is extremely important to any researcher. This led us to the research question at hand: "Can measurable features of a scholarly paper reliably classify it as belonging to high impact or low impact journals?"

This is a relevant issue particularly for graduate students (rejection delays graduation), early-career researchers (unclear journal targeting), faculty members looking for promotion (wasted time due to mismatched journal submissions) and universities or research institutes (uneven publication performance across departments).

To answer the research question, this study utilizes the predictive analysis and machine learning techniques. Structural paper characteristics are used, like page count, word count, citation intensity, author count, visual density, paper type and field domain. As Impact factor is widely used as a main criteria for assessing research quality, so it serves as the best target variable for this classification task. Findings aim provide insights to researchers based on evidence for journal targeting strategy and publication.

Our success criteria reflect the real-world costs of journal targeting decisions. Since incorrectly predicting a paper as suitable for a high impact journal leads to lost time and delayed publications, our evaluation emphasizes precision and overall model reliability.

We hope that researchers and institutions will be able to utilize our model to avoid waste and improve publication outcomes. For researchers, this will be done by tailoring their writing to improve their competitiveness. This doesn't necessarily mean that individuals should completely change their research to achieve high impact publication, but instead guide them to appropriate journal submissions. As for institutions, our findings can support mentoring programs and help departments better guide early career researchers through the publication process. This can ultimately strengthen overall research visibility and improve institutional standing.

Literature Review

According to Eugene Garfield (2006), the impact factor measures the average number of citations received by articles published in a journal over a specific period of time. Initially, it was developed to compare journals on the basis of citation performance, and it later became a widely used criterion for selecting appropriate journals for research publication.

However, citation-based indicators are influenced by multiple factors such as discipline, journal reputation, publication type, and scholarly behavior.

Iman Tahamtan and Lutz Bornmann (2019) highlighted that citation counts depend on several variables including paper quality, author reputation, collaboration, visibility, number of authors, and journal impact factor. Their findings suggest that journal impact is not determined by a single feature but rather by a combination of structural, contextual, and

scholarly characteristics.Previous studies have also shown that article characteristics such as length, collaboration, and references significantly influence research visibility.

Matthew E. Falagas et al. (2013) found that longer articles are more likely to receive higher citation counts because they generally contain detailed content and greater scientific complexity.

Stefano Mammola et al. (2021) reported that highly cited papers often include a larger number of references and cite recent as well as influential scholarly sources. These findings support the inclusion of variables such as word count, page count, and number of references as important predictors in publication impact analysis.The importance of research collaboration has also been emphasized in prior literature.

Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi (2007) observed that multi-author research has become increasingly dominant across academic disciplines, and collaborative work generally receives more citations than single-author studies. This supports the use of author count as a predictor variable because collaborative papers may benefit from diverse expertise, stronger academic networks, and greater research capacity.

In addition, article type and field domain also influence citation behavior and publication outcomes. Review articles, empirical studies, conceptual papers, and commentary articles differ considerably in terms of reference count, article length, and citation patterns. Citation norms also vary significantly across disciplines such as business, healthcare, social sciences, and artificial intelligence.

Machine learning techniques are increasingly being used in predictive analytics within scholarly communication research. Previous studies on citation prediction have utilized paper metadata, semantic features, bibliometric indicators, and citation behavior to estimate future research influence.

Atefeh Abrishami and Saeed Aliakbary (2019) developed a neural network model for predicting citation counts. More recent research has incorporated metadata and semantic information to improve prediction accuracy for research impact. Although these studies demonstrate the usefulness of predictive analytics in evaluating research potential, most of them focus on post-publication citation prediction rather than assisting researchers in journal selection before manuscript submission.

Existing literature indicates that factors such as impact factor, references, article length, collaboration, and metadata may influence research visibility and scholarly impact. However, very few studies have specifically examined whether measurable structural characteristics of academic papers can classify them as suitable for high-impact or low-impact journals prior to submission.

Research Gap

There is much literature available on citation prediction and impact analysis of research but there is limited research available on whether measurable structural characteristics of paper can predict or classify the paper as potentially low impact or high impact journal publication.

Most studies focus on citation accumulation post publication rather than on journal selection strategy during research work submission preparation. This study addresses this gap through predictive analytics tools to classify the papers as belonging to either high-impact or low-impact journal using measurable features before submission. By staying focused on publication strategy - this study contributes to both managerial and theoretical insights for researchers, institutions and academic development.

Methodology

Research Design

This study uses predictive analytics to classify a paper to be potentially published in a high-impact and low-impact journal based on paper's structural characteristics. Machine-learning classification models were used to identify the relationships between features and journal impact.

Data Collection

The dataset – 104 academic papers were collected from multiple scholarly databases such as ScienceDirect, IEEE, Emerald, JMIR, tandfonline and MDPI. The papers selected to be used in our analysis fell within the following domains: Business, Healthcare, and Social Sciences, with some papers having a combination of one of these domains alongside AI.
To ensure data consistency, we manually extracted variables using standardized data collection method. Page, figure and author counts, references and paper types were recorded uniformly across all observations.

Variables

Several variables in our dataset exhibit natural outliers and skewed distributions. Impact factor values are right skewed due to a small number of exceptionally high impact journals. Page count, word count, and number of references also show right skewed patterns, with review papers in particular affecting word counts. These distributions are expected and will be accounted for during modeling and evaluation.

Table 1 - Summary of Variables Used in the Analysis

Variable	Description	Data Type	Relevance
Impact	Journal prestige rating	Continuous	Target classification variable

Factor			
Page Count	Total pages in paper	Integer	Reflects paper depth
Word Count	Total estimated words	Integer	Indicates comprehensiveness
Number of Authors	Total authors listed	Integer	Reflects collaboration intensity
Number of Figures	Tables and figures combined	Integer	Measures visual presentation
Paper Type	Review, experimental, conceptual, commentary	Categorical	Captures publication format
Number of References	Total cited references	Integer	Indicates theoretical depth

Feature Engineering

Here, we created new variables that might help the model understand the papers better**.** For example, Visual density was calculated by dividing the number of figures by number of pages. This shows how "visually heavy" a paper is. Citation intensity, which tells us how many references a paper has for every 1000 words.

Target Variable Transformation

Impact factor values were transformed into a binary classification: High (impact factor greater than 7) and Low (impact factor less than or equal to 7). This cutoff came from what many people in the business journal world consider "high impact."

Data Cleaning and Transformation

One of the observations had zero references, it was removed because logarithmic transformation produced infinite value. We also checked that all variables had the correct data types. For outliers in references, we applied log to it.

We also handled the categorical variables by turning the field of the paper into dummy variables and turning the impact factor into High/Low groups. These transformations helped the model understand categories in a simple way.

Model Selection

Four classification models were evaluated:

1. Generalized Linear Model (GLM)
2. Logistic Regression
3. Decision Tree
4. Random Forest

These models were selected because of interpretability, ability to classify and practical applicability.

Evaluation Metrics

Model performance was evaluated based on accuracy, specificity, area under the curve (AUC), confusion matrix analysis and lift chart analysis. The study underlined specificity due to high cost associated with false positive predictions in targeting journals.

Exploratory Data Analysis

Exploratory data analysis was done to identify patterns, assess data quality, and determine which variables may be most informative for classification.

Table 2 – Descriptive Statistics of Numerical Variables

Variable	Mean	SD	Min	Max
Impact Factor	5.538	3.637	0	21.6
Number of Figures	8.567	6.148	0	29
Page Count	18.721	9.871	6	50
Word Count	9473.279	5811.661	1300	49000
Author Count	3.750	2.783	1	17
Number of References	77.798	48.719	0	270

The descriptive statistics represents substantial variation across the paper characteristics. Impact factor demonstrated right-skewness due to several extremely high-impact journals and a similar skewness patterns is visible in references, word count, and figure variables.

Class Distribution

The target variable distribution revealed moderate class imbalance – 75 low-impact papers and 29 high-impact papers suggesting that model evaluation meticulously consider classification bias toward the larger class.

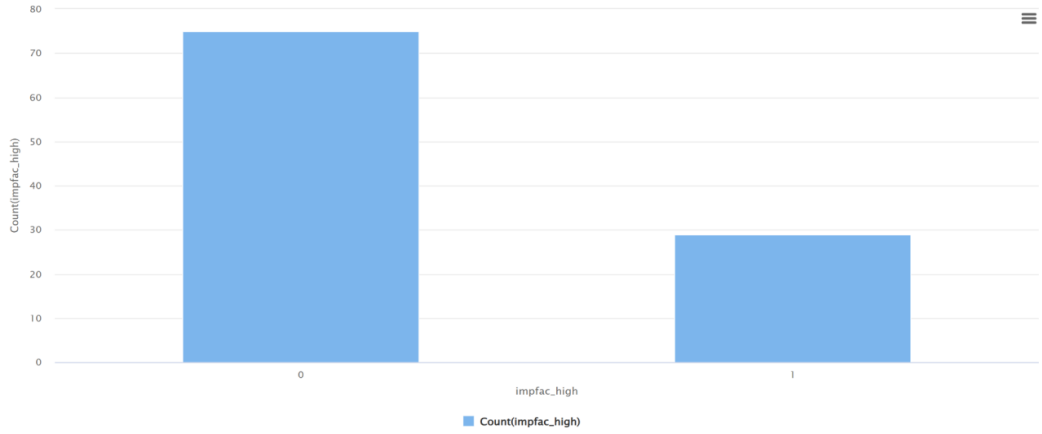

Figure 1: Bar plot of target class frequency (0=low impact, 1=high impact)

Correlation Analysis

It revealed moderate to strong relationships among variables like page count, word count, figures, and references. Longer papers mostly had more references and visual elements. On the other hand, impact factor shows only very weak correlations with all numerical features, with values close to zero. This suggests that the simple linear relationships are not sufficient to distinguish high impact from low impact journals and reinforces the need for predictive classification modeling.

Attribu...	figures	pages	words	authors	referen...	impfac...
figures	1	0.400	0.540	0.036	0.211	0.007
pages	0.400	1	0.629	-0.096	0.387	0.029
words	0.540	0.629	1	-0.023	0.547	0.077
authors	0.036	-0.096	-0.023	1	0.186	0.121
references	0.211	0.387	0.547	0.186	1	0.165
impfactor	0.007	0.029	0.077	0.121	0.165	1

Figure 2: Correlation Matrix

Boxplot Analysis

High-impact papers generally demonstrated higher word and page counts, more references and higher visual density. Although overlap existed between two groups, the overall trends illustrates that structurally comprehensive papers are more likely to be published in high-impact journals.

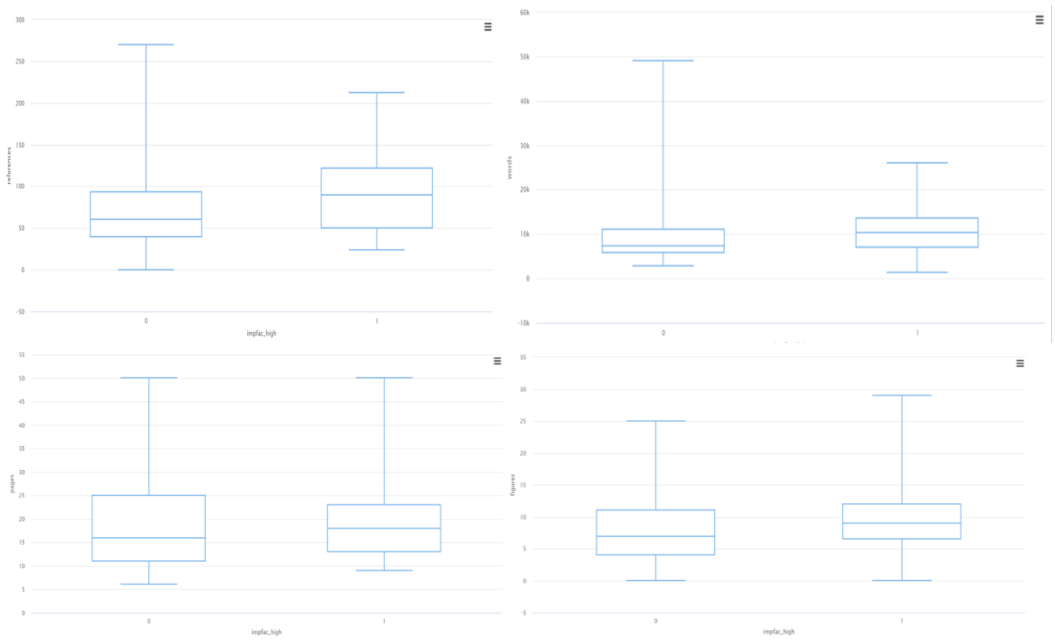

Figure 3: Boxplots of References(top L), Words (top R), Pages (bottom L), Figures (bottom R) w.r.t. Impact

Results

Model Performance

Logistic regression demonstrates the best accuracy = 0.70 i.e. 7 out of 10 cases were correctly classified with a specificity = 0.96 i.e. model's ability to identify papers correctly which are likely to achieve success or not based on metrics. Though Generalized linear model has highest AUC=0.68, the best model would be logistic regression based overall performance.

Table 3 – Classification Model Performance

Model	Accuracy	Specificity	SD	AUC
Logistic Regression	0.70	0.96	0.09	0.65
Generalized Linear Model	0.47	0.39	0.17	0.68
Random Forest	0.60	0.71	0.08	0.61
Decision Tree	0.60	0.86	0.13	0.50

Confusion Matrix Analysis

The confusion matrix demonstrated 21 true low predictions, 8 false negatives, 1 false positive, 0 true high predictions. Model effectively screened the low-impact papers while it is less reliable in identifying high-impact papers. The extremely low false positive rate is significant because incorrect recommendation for potentially high-impact journal paper may result in delays in publication and rejection risk. However, the model struggled to accurately identify true high-impact papers – zero sensitivity; likely resulted from class imbalance, small dataset and overlapping feature distributions.

	true Low	true High	class precision
pred. Low	21	8	72.41%
pred. High	1	0	0.00%
class recall	95.45%	0.00%	

Figure 4: Confusion Matrix

Lift Chart Analysis

Analysis on model probabilities shows strong specificity which means that high proportion of correctly classified low-impact papers forms the top deciles. the lift curve exhibit the utility to screening out weak papers, but the limitation is it's ability to identify highly successful ones; lift curve shows limited separattion between both.

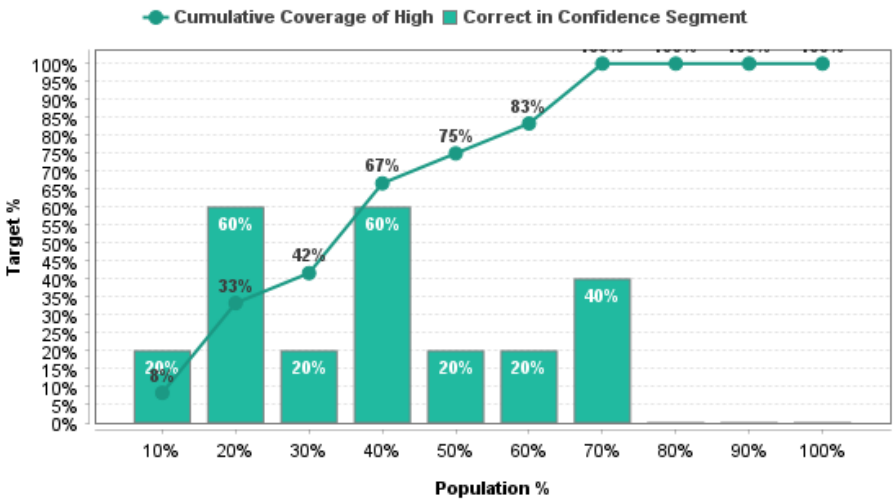

Figure 5: Lift Chart – Logistic Regression

ROC Curve Analysis

The logistic regression model achieved an AUC=0.65, indicating moderate discrimination capability. While logistic regression model performed better than any other model, the results suggest limited predictive ability in classifying high-impact and low-impact papers.

The ROC analysis reinforces the findings that model is better suited for screening out potential low impact papers rather than successfully identifying high impact papers.

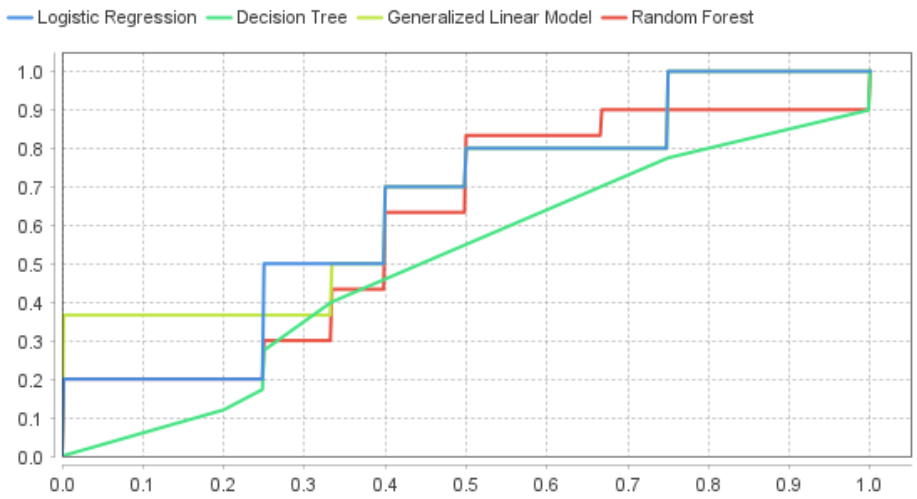

Figure 6: ROC Curve analysis

Feature Importance and Interpretability

Many features represented strong relationships with impact factor classification:

- Citation intensity strongly correlates with impact level.
- Visual density contributes positively to publication success.
- Lengthy papers demonstrated higher probability of belonging to high-impact journals.
- AI domain papers demonstrate higher likelihood of becoming high-impact.

These findings indicate that structurally rich, visually supported and theoretically grounded papers may align better to requirements of prestigious journals.

Discussion

Findings suggests that measurable features of a research paper can partially predict the journal classification – impact factor. While models show limited sensitivity, few important trends appeared which provide insights for researchers and research institutions.

Citation intensity was one of the strongest predictors, papers with high scholarly engagement are more likely fit for high-impact journals; which also supports prior literature focusing on the importance of literature integration and theoretical depth.

Visual density also led to positive publication outcomes. Figures, tables and visual representations may enhance clarity, reader engagement and methodological transparency specifically within AI, healthcare and business research.

Lengthy paper demonstrates a positive relationship with publication impact. High-impact journals may prefer comprehensive analyses which provides deeper details and stronger contributions.

Additionally, more authors slightly increase publication odds and more references have a small positive effect as well. Writing a research paper alone is not the best idea.

The stronger performance of logistic regression highlights that interpretable machine learning is important for scholarly analytics. Unlike the more complex – black-box models, the logistic regression allows researchers in understanding how individual features of paper can influence classification outcomes.

However, the present study reveals significant limitations in identification of high-impact papers. The model's low sensitivity also suggests that only structural features cannot explain publication success. Qualitative factors like novelty, writing quality, reviewer preferences, institutional prestige and author reputation also play significant role.

Recommendations

- **Recommendations for Researchers:** Prioritizing stronger literature integration, visual clarity and comprehensive content development may benefit as it provides deeper analysis and detailed methodology.
- **Recommendations for Institutions:** Universities should incorporate the publication metrics into research and faculty development programs. For early researches, predictive tools may help in improving journal targeting strategies and optimize submission cycles.
- Predictive analytics can be used as a decision-support tool and not as a decision making or recommendation system.

Limitations

The model cannot yet identify papers with a high impact if its sensitivity is zero. Also, dataset which is 103 papers, is small and imbalanced and covers only few high-impact cases. Information on some of the impactful features like quality of abstract, author's reputation, institutional ranking or cite score were not available for all the papers. Since target class is uncommon, it contributed to the 0 true-positive. Only structural characteristics may not completely capture complex peer review and editorial processes which influences publication success.

Future Research

Future research in this area should expand the dataset by adding more disciplines, publication years and journal categories. Larger datasets would help in improving model stability and also support more advanced machine learning techniques.
Future studies may incorporate scholarly influence features such as h-index, citations per paper, funding data or metrics of collaboration network.
Balancing the dataset through oversampling or penalized classification techniques may also improve sensitivity and high-impact detection capability.

Conclusion

This study determines whether measurable structural characteristics of academic papers can classify the paper as either potentially high-impact or low-impact journal publication. Predictive analytics tools and machine learning techniques helped in analysis of features like citation intensity, visual density, field domain and paper length which shows partial contribution in classifying publication impact.
Logistic regression model demonstrated strongest overall performance with better accuracy, specificity and interpretability as compared to other models. While model struggled in identifying true high-impact papers; the findings suggest that predictive analytics could support the evidence-based strategies for publication and improve decision making for targeting journals.
The results highlighted the increasing potential – Integrating machine learning into scholarly publications. Although predictive analytics tools should not replace the expert judgment, but they might serve as a valuable tool for decision-support optimizing submissions and publication.
Overall, optimizing publication strategies needs a balance between quantitative insights and qualitative scholarly judgment. As analytics in research continues to evolve, the data-driven approaches could become increasingly crucial and help researchers in navigating competitive publishing in academic landscape.

References

- Garfield E. (2006) The History and Meaning of the Journal Impact Factor. JAMA. 2006;295(1):90–93. doi:10.1001/jama.295.1.90
- Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated

review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics, 121*(3), 1635–1684. https://doi.org/10.1007/s11192-019-03243-4

- Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals. *PLOS ONE, 8*(2), e49476. https://doi.org/10.1371/journal.pone.0049476
- Mammola, S., Piano, E., Doretto, A., Caprio, E., Chamberlain, D., & Isaia, M. (2021). Measuring the influence of non-scientific features on citations. Scientometrics, 126, 7745–7760. https://doi.org/10.1007/s11192-020-03759-0
- Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099
- Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. Journal of Informetrics, 13(2), 485–499. https://doi.org/10.1016/j.joi.2019.02.011