Xu, Huimin The University of Texas at Austin, USA | huiminxu@utexas.edu
Saar-Tsechansky, Maytal The University of Texas at Austin, USA | Maytal.Saar-Tsechansky@mccombs.utexas.edu
Song, Min Yonsei University, South Korea | min.song@yonsei.ac.kr
Ding, Ying The University of Texas at Austin, USA | ying.ding@ischool.utexas.edu

ABSTRACT

citations thro ugh team-related variables, team composition, and team structure. Team composition includes team size, male/female dominance, academia/industry collaboration, unique race number, and unique country number. Team structures are made up of team

citation number, H-index, previous collaborators, career age, and previous paper numbers are a proxy of team power. We calculated the mean value and Gini coefficient to represent team power level (the collective team capability) and team power hierarchy (the vertical difference of power distribution within a team). Taking 1,675,035 CS teams in the DBLP dataset, we trained the XGBoost model to predict high/low citation. Our model has reached 0.71 i

importance in predicting team citation categories, we found that team structure plays a more critical role than team composition in predicting team citation. High team power level, flat team power structure, diverse race background, large team, collaboration with industry, and male-dominated teams can bring higher team citations. Our project can provide insights into how to form the best scientific teams and maximize team impact from team composition and team structure.

KEYWORDS

Citation predication; Team composition; Team structure; XGBoost; Explainable AI; SHAP

INTRODUCTION

Citing behavior represents the credits to those scientific works with intellectual contributions (Merton, 1988). Yet some researchers argue that high citations do not mean disruptive innovation to the scientific community (Wu et al., 2019), and citations bring structural biases to minority groups, disparities across institutions and countries (Sugimoto, 2021). Despite these shortcomings, the citation count is still a crucial metric to evaluate the impact of scientific work, individual scientists, journals, and departments (Bornmann et al., 2008). For example, the low citation of papers will influence how people perceive their quality and thus reduce close reading (Teplitskiy et al.,

ranking (Cole & Cole, 1967). The citation of a paper also quantifies the impact of a scientific team who collaborates to finish the publication. Given the citation of papers requires the accumulation of time, it is a valuable task to predict the long-term team impact shortly after the paper's publication.

Scientific work is more team collaboration than individual endeavors (Wuchty et al., 2007). Division of labor requires everyone to specialize in specific areas to solve complex problems (Becker & Murphy, 1992). Previous research predicts paper citations from paper features, journal features, author features, early citation, and peer review (e.g., Li et al., 2019; Ruan et al., 2020). But, the literature shows poor research endeavors in predicting citation count regarding team-related features. The science of science research community mainly focuses on exploring and stressing individuals' characteristics to influence team impact instead of taking all researchers in the team as a whole. Specifically, Li et al. (2019) measured the characteristics of first authors (e.g., the accumulated citations, H-index, publication number) in the author features but ignored the contributions of all other authors. Compared with the influence of an individual, the team structure of how members communicate (Woolley et al., 2010), share power (Xu et al., 2022a) and divide the work (Xu et al., 2022b) has a greater influence on team performance. Besides the characteristics of the first authors or corresponding authors, the team composition, such as team size (Wu et al., 2019), gender and race diversity (Hofstra et al., 2020), country diversity (Wagner et al., 2019) of team members are also determinants of team success. In this project, we aim to predict team impact using team-related variables, team composition, and team structure. Teams in this project is represented by the list of authors of one published article because the authors of one publication have to work together as a team for one or two years to get this paper published. As to team composition, we consider team size, racial backgrounds, institutional backgrounds, academia/industry collaboration, and gender-dominated variables for all authors. Compared with team composition, team structure is widely applied in organizational behaviors (Greer et al., 2010) and sports (Halevy et al., 2012). Still, few are discussed in scientific teams (Xu et al., 2022). Team structure comprises team power level and team power hierarchy. Team power level means the collective ability of the whole team (Greer et al., 2011). Team power

hierarchy implies the distribution of resources within the teams (Greer et al., 2018). It can be hierarchical (e.g., top-down management) or flat (all team members share similar power), influencing the communication, collaboration process, and performance. Team power can be measured using team members' previous citation number, H-index, previous collaborators, career age, and previous paper numbers. We can compute the average value and Gini coefficient to represent team power level and hierarchy.

Integrating prediction and explanation is essential when we study social problems to understand how and why the phenomenon happened (Hofman et al., 2021). The explanation can help us open the black box of machine learning models. Instead of pursuing high accuracy, we also need to know models' ethical implications and consequences (Wallach, 2018). For example, although a model has high accuracy, it might bring bias to disadvantaged people. However, the research on paper citation prediction lacks a combination of high prediction and robust explanation. There are two branches of study in predicting team citation. One branch focuses on achieving high prediction results but lacks a thorough analysis of how variables contribute to final results. The other branch highlights the interpretability of models but lacks high predictability of team performance. In this paper, we choose XGBoost (Chen & Guestrin, 2016) to help predict team impact measured by citations. By applying the Explainable AI method SHAP (Lundberg & Lee, 2017), we can quantify feature contribution to the citation count. Meanwhile, we can understand how team features influence each team locally and all teams globally through interpretability. Based on that, we can provide insights into team formation in the scientific context.

This paper uses explainable AI methods to understand team formation and team impact. This study addresses the above research gaps by taking Computer Science (CS) as a test field to predict team citation through team structure and composition variables. This paper is structured as the followings. Section 2 summarizes the related literature. Section 3 details the research methods. Section 4 describes the model comparisons and feature contribution. The last Section 5 concludes the study and points out future research directions.

RELATED WORK

Features in Citation Prediction

The literature mainly predicts paper citation based on five aspects of information: 1) paper content, 2) peer review, 3) journal, 4) early citation, and 5) authors. Firstly, the paper's content, such as title, abstract, keywords, topic, and references, is used to predict paper citation count (Fu & Aliferis, 2010; Hu et al., 2020; Jiang et al., 2021, Yan et al., 2011). For example, Hu et al. (2020) applied the LDA model to extract keywords from titles, abstracts, and author-defined keywords. They then measured these keywords popularity on google scholar, research gate, and google trend websites. They suggested that keywords popularity was an important indicator of citations. Also, when papers cite the most recent and relevant references in the subfield, they would attract more attention in the community (Roth et al., 2012). Secondly, peer review, as an evaluation of paper quality, Li et al. (2019) extracted the most similar peer review texts to abstracts and proposed a cross-review matching mechanism in order to get a comprehensive and relevant review representation. Thirdly, journal quality, like, the impact factor, is an important predictive feature for paper citation (Fu & Aliferis, 2010; Yan et al., 2011). Fourthly, early citations of papers can contribute to long-term citations (Newman, 2014). Researchers predict 5-year citations based on the first two years (Ruan et al., 2020). The final aspect of author information is the main element in predicting citation. Researchers select the first, most productive, and most influential authors to help predict the whole team's performance (Fu & Aliferis, 2010). Usually, researchers might use a mixture of this content-based and bibliographic information to predict paper impact jointly. This paper uses a unique perspective team-related variables to predict paper citation. We take all individuals in the paper as a team to predict team outcomes.

Team Composition and Structure

Team composition mainly highlights demographic-related variables influencing team performance, such as team size, gender, race, sector, and country. For example, by analyzing the team photo after finishing the game, Saveski et al. (2021) trained machines learning to judge whether the team escaped successfully from the maze. They found that large teams, gender diversity, race diversity, and older but fewer age differences can improve the chance of success. Team size plays an important role in team performance. Team size, as a parameter in a model for the self-assembly of creative teams, could determine team performance (Guimera et al., 2015). Large teams tend to receive more citations (Wuchty et al., 2007; Larivière et al., 2014). However, Wu et al. (2019) showed that innovation does not scale up with large teams. As to gender composition, researchers designed games to illustrate how the proportion of different gender influences the process of cooperation in lab experiments. They concluded that female-dominated teams are positively related to equal conversations (Woolley et al., 2008). This is because women have more social sensitivity compared with men in teams, which means they can sense how others think through nonverbal signals. This result is consistent with an experiment that requires students to finish a project paper in class (Berdahl & Anderson, 2005). They found that majority-female and balanced teams prefer equal communication and shared leadership over time, which could encourage everyone to participate in activities and contribute to teams. Freeman & Huang (2015) studied how race diversity influences team performance in the WOS dataset. They divided race into eight categories and found that although scientists prefer to work with those who are similar to them in race, homophily can lead to lower citations and low-impact journals. Thus, they confirmed an assumption race diversity has positive effects on teams. Compared with the industry, universities dominate in paper publishing (Larivière et al., 2018). The collaboration between these two different sectors can bring more benefits for teams. For example, by analyzing the paper published by Canadian institutions from 1980 to 2005, Lebeau et al. (2008) confirmed that university-industry teams have more citations than pure university or industry teams. When authors from different countries are involved in teams, it can cause an additive citation effect (Hsiehchen & Hsieh, 2015). Wagner & Leydesdorff (2005) also proved that international collaboration could receive more citations for papers. Overall, the diversity of demography in teams can increase team citation.

The team structure is made up of team power level and team power hierarchy. The concept of team power level and team power hierarchy has been discussed in different contexts. In the organizational behavior field, Greer et al. (2010) regarded the positions in organizations as power and computed the average positions as team power level and standard deviation as team power hierarchy. In sports, Halevy et al. (2012) measured team power with NBA players' salaries. They all found that team structure is crucial to team conflicts, team coordination, and team ultimate performance. When people have brainstorming, negotiating, and puzzle games in experiments, a flat structure measured by equal speaking-turn distribution among team members can maximize collective intelligence (Woolley et al., 2010). In the scientific context, as long as more than one person works on a common project, team power dynamics will influence every step of collaboration, including defining topics, distributing labor of division, choosing methods, deciding journal targets, etc. A researcher's career age, previous citation, H-index, previous collaborators, productivity, function, and role can also be taken as the proxy of power since they represent prestige, resources, knowledge, and experience (Merton, 1973). Xu et al. (2022a) took the career age as the basis of power, where senior researchers have more rights in decision-making than junior researchers. By analyzing teams in Computer Science, Physics, Sociology, Library & Information Science, and Arts & Humanities, Xu et al. (2022a) calculated the mean value and Gini coefficient of team members' career age to represent the team power level and team power hierarchy. Finally, they found that a flat structure at different team power levels can have higher team citations. Similarly, Xu et al. (2022b) identified the leaders who conceived, designed, supervised, and wrote in publications through paper contributions and suggested that a flat structure with multiple leaders was better for team novelty and long-term impact than sole leadership. Although these team-related features have been proved important to performance, few science of science research considers them when predicting citation.

Methods in Citation Prediction

Predicting paper citations uses either traditional statistics models or advanced deep learning models. Yu et al. (2014) applied multiple linear stepwise regression models to analyze the correlation between paper features and citations in around 1,000 papers in the field of Information Science & Library Science. Similarly, Fu & Aliferis (2010) used SVMs and logistic regression models to predict whether medical papers exceeded a threshold value within ten years. These simple models are easily interpret-ed by understanding the coefficient of related variables. Ma et al. (2021) utilized deep learning models Bi-directional Long Short Term Memory (Bi-LSTM) to predict AI paper citations in the long term. Even though they found deep learning models out-perform other baseline models and achieve higher accuracy, they did not interpret the outputs from the black box model. In the library, information and document field, Ruan et al. (2020) predicted papers' citations within five years through the BP neural network. To explain the relative importance of features, they trained different models when they dropped one feature but kept all other features (leave-one-out model) and then compared model differences. The drawback is authors do not consider the correla-tion between variables, and the model computation has low efficiency. Integrating prediction and explanation in modeling is a future direction since explanatory power can quantify how much we explain and reveal the limitation of our understanding of the social phenomenon (Hofman et al., 2021).

The combination of XGBoost and SHAP has been used in various contexts, such as marketing, health care, and traffic safety. By analyzing the product reviews on Ama-zon with XGboost and SHAP, Meng et al. (2020) found that the length of reviews contributes most to the headset review helpfulness while the frequency of product attributes is most helpful to the facial cleanser consumers. Yang (2020) trained the XGboost model to predict patients' recovery and modality. Compared with gender, time in the hospital, and the presence of chronic disease, Yang (2020) concluded that age is the most important factor in COVID-19 mortality prediction through SHAP and LIME value contributions. In accident safety, when inputting traffic, network, demographic, land use, and weather features into the XGBoost model, the speed dif-ference before and after accidents is the key to traffic occurrence (Parsa et al., 2020). The accuracy can achieve 99% in the Chicago metropolitan highway dataset. Ma et al. (2022) predicted the visibility of papers that are broadly mentioned and long-term disseminated on social media through literature-related features (authors, journal, paper) and social media-related features (user, tweets). They found that XGBoost algo-rithms performed best in predicting paper visibility and explained the contributions of each feature in disseminating through the SHAP method. Overall, the methods can achieve a balance between a high prediction rate and high interpretability. But the explanation AI method has not been widely applied in predicting and understanding team citation.

METHODOLOGY

Data

We use the DBLP dataset (https://dblp.org/), including 4,894,081 papers in the Computer Science (CS) field. We choose 3,658,127 papers that have more than one author from 1980 to 2020 since we assume a team must have at least two authors. We take authors in each paper as an independent team who collaborate together to finish a research project and publish the outcomes. For example, if a paper has five authors, then these five authors are considered to be a team. When we identify the gender and academia/company information of each author, there are unknown categories. Unknown gender-dominated teams occupy 27% (984,912) and unknown race teams occupy 36% (1,316,527). We exclude these unknown teams, and 1,675,035 teams are left with complete information.

Measures

Team impact: We use the 5-year citation to evaluate the impact of a paper, which is also the impact of a team who published this paper. This is a binary variable, high team impact (above or equal to mean value) and low team impact (below mean value). The mean 5-year citation is 5. 1,263,162 teams are below the mean value, whereas 411,873 teams are above the mean value. Above-mean value is labeled as 1, and the below-mean value is labeled as 0. Here, we do not use the median value of a 5-year citation as a threshold. The median value of a 5-year citation is 2 since most of the papers (63%) do not have more than two citations. Using the median value cannot show the distinct difference between high-impact papers and low-impact papers.

Team size: We count the number of authors in a team as the size of this team. The median team size value is 3.

Gender: Teams are assigned to male-dominated, female-dominated, and equal. For instance, male-dominated means the identified male author number is larger than the identified female author number. We use the Bert-based model trained by Acuna & Liang (2021) to predict authors' gender information. After excluding unknown categories, there are 1,316,193 (79%) male-dominated teams, 184,608 (11%) equal teams and 174,234 (10%) female-dominated teams.

Race: We calculate the team's unique number of authors' race. Acuna & Liang (2021) distinguished four kinds of race: Black, Hispanic, White, and Asian. The median value of unique race in teams is 1, suggesting that most of the teams are homogenous in race.

Sector: Teams are divided into pure academia, pure company, and combined. Pure academia (company) means teams are made up of all researchers from academia (industry) institutions, whereas combined teams have both academia and industry researchers. Manjunath et al. (2021) methods have been applied to match authors' institutions with eight categories in the Global Research Identifier Database (GRID), including government, education, company, facility, healthcare, nonprofit, archive, and others. We merge the health care (mainly universities) and education into academia and keep the company as the industry category. After excluding unknown categories, there are 1,501,131 (90%) academia teams, 63,399 (4%) industry teams, and 110,505 (7%) combined teams. Academia teams are dominating in paper publishing.

Country: We calculate the unique number of authors' countries based on authors' affiliations in a team. The mean value of a unique country in teams is 1, indicating that most teams come from the same country.

Career Age: We calculate the mean career age and Gini index of career age among all team members. The median value of mean career age is 8, and the Gini career age is 0.34.

Citation: We calculate the mean citation and Gini index of citation among all team members before collaborating the paper. The median value of mean citation is 63, and the Gini career age is 0.50.

H-index: We calculate the mean H-index and Gini index of the H-index among all team members. H-index can comprehensively reflect a researcher's productivity and impact (Hirsch, 2005). The median value of the mean H-index is 3, and the Gini H-index is 0.50.

Collaborators: We calculate the mean collaborators and Gini index of collaborators among all team members before collaborating on the focal paper. The median value of the mean collaborator is 19.5, and the Gini coefficient of collaborators is 0.39.

Productivity: We calculate the mean papers and Gini index of papers among all team members before collaborating on the focal paper. The median value of mean career age is 15.25, and the Gini career age is 0.50.

Methods

An XGBoost model is trained for binary classification of high/low team impact. XGBoost applies a gradient tree boosting algorithm and has state-of-the-art classification performance (Chen & Guestrin, 2016). Firstly, we plan to use Python xgboost package. 80% dataset was taken as training, and the remaining 20% dataset was taken as testing. For the predicted value, the result is not balanced. Below-mean values have a more significant proportion. Thus, we take the up-sampling methods to ensure the trained dataset is balanced. We randomly choose from the positive values multiple times to make the positive instances number the same as the negative ones. Secondly, we evaluate the model's performance in comparison with baselines. Thirdly, SHAP is an explainable AI method used for model interpretation through Shapley value (Lundberg & Lee, 2017). By comparing the difference with the variable and without the variable in the model, the importance of each variable can be inferred. SHAP interprets the contribution values of each feature in each instance to the particular prediction. We can draw the relative feature importance of explanatory variables on the model results using the mean SHAP value of variables and the distribution of the SHAP value of variables for each individual. Finally, we draw interactive plots to show the relationship between variables and SHAP values.

Evaluation Metrics

The results of the predicted model can be divided into four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Here 0.5 is used to be a threshold in the 2-binary classification where the instance with value larger than 0.5 is a positive one, and the instance with a value lower than 0.5 is a negative one. We calculate the accuracy (ACC) as the evaluation metrics: ACC = (TP+TN)/(TP+FP+TN+FN) . F1 measure combines the consideration of Precision, Precision = TP/(TP+FP), and Recall rate, Recall = TP/(TP+FN), F1= (Recall*Precision*2)/(Recall+Precision). Difference from using 0.5 as threshold values, AUC considers the performance of different thresholds. AUC considers the variables of Recall, and False Positive Rate, FPR =

- 1.

Hyper-Parameter Tuning

In the XGBoost model, we set the learning rate is 0.01, the subsample ratio of the training instances is 0.8, the maximum depth of a tree is 20, the minimum sum of instance weight in a child is 5, the minimum loss reduction required to make a further partition on a leaf node of the tree is 3. Fig 1 shows the results of Trian and Test within 100 epochs. Our objective of the model is to maximize the AUC and Error (1-ACC) value. It suggests that the changes in error and AUC values in the test dataset have been stable. Thus, we choose the last epoch as the best performance of our model.

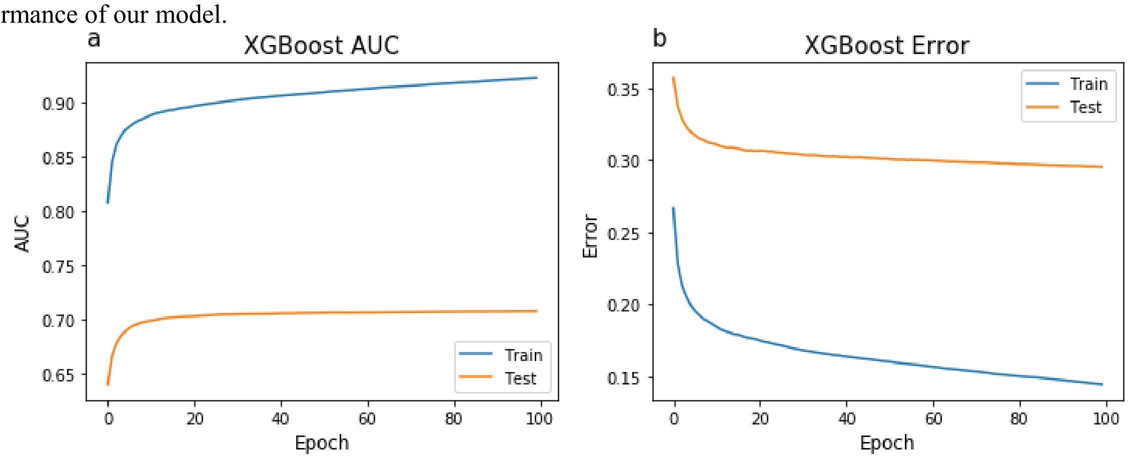

Figure 1. Evaluating accuracy results of XGBoost Model on DBLP dataset

RESULT

Comparison of Model Performance

We choose logistic regression and decision tree as our baseline models. We can conclude that the XGBoost model outperforms these traditional models in all three evaluation metrics: ACC, F1, and AUC. Although the XGBoost model only has a 4% higher accuracy rate than the logistic model, we need to consider that our sample is large, so the 4% improvement means a large sample.

Model	ACC(%)	F1(%)	AUC
XGBoost	70.45%	46.84%	0.71
Logistic Regression	66.38%	45.76%	0.68
Decision Tree	62.72%	44.72%	0.66

Table 1. Model comparison results on DBLP dataset

Results of Feature Contribution

When we choose to predict models, we can feed more features into the model to increase the prediction performance. When we explain how variables contribute to prediction performance, we need to consider the interdependency between features, especially the team structure variables, such as career age, H-index, productivity, citation, and collaborator. These team structure variables are highly related (Table 2-3). We can find that the H-index is highly related to career age, productivity, citation, and collaborator, in the combination of mean value and Gini coefficient. Thus, when we try to explain the feature contribution, we only consider H-index in the team structure.

	Mean Career Age	Mean H-index	Mean Productivity	Mean Citation	Mean Collaborator
Mean Career Age	1.00	0.69	0.61	0.42	0.54
Mean H-index	0.69	1.00	0.84	0.82	0.77
Mean Productivity	0.61	0.84	1.00	0.71	0.87
Mean Citation	0.42	0.82	0.71	1.00	0.64
Mean Collaborator	0.54	0.77	0.87	0.64	1.00

Table 2. Pearson correlations between mean team structure variables

	Gini Career Age	Gini H-index	Gini Productivity	Gini Citation	Gini Collaborator
Gini Career Age	1.00	0.74	0.74	0.66	0.68
Gini H-index	0.74	1.00	0.84	0.87	0.51
Gini Productivity	0.74	0.84	1.00	0.84	0.63
Gini Citation	0.66	0.87	0.84	1.00	0.54
Gini Collaborator	0.68	0.51	0.63	0.54	1.00

Table 3. Pearson correlations between gini team structure variables

Cases

We visualize two cases (Fig 2) to show how features contribute to each team locally. Comparing the first team with the second one, we can find that lower Gini H-index, higher Mean H-index, and larger team size positively contribute to team citation.

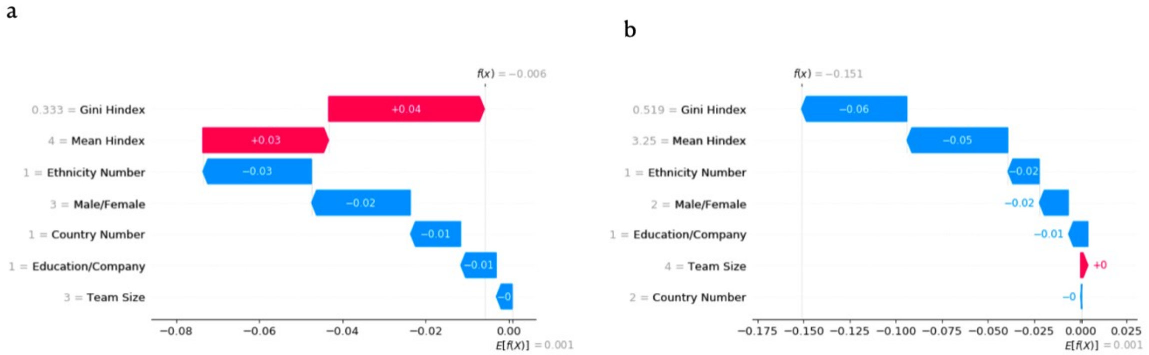

Figure 2. SHAP feature contribution of two teams

SHAP Feature Importance

We apply the SHAP method (Lundberg & Lee, 2017) to better interpret the XGBoost model's results. This method measures the contribution of each feature to each team in the DBLP dataset and then aggregates the results of all samples. The aggregated mean (|SHAP value|) are ranked from high to low. In Fig 3, team power level (Mean H-index) and team power hierarchy (Gini H-index) are the critical factors, followed by Race Number, Team Size, Academia/Industry, Male/Female, and Country Number. The mean H-index of the whole team members stands out in the first position, and meanwhile, the distribution of the H-index within the whole team is important. We can conclude that team structure variables are more important than team composition variables.

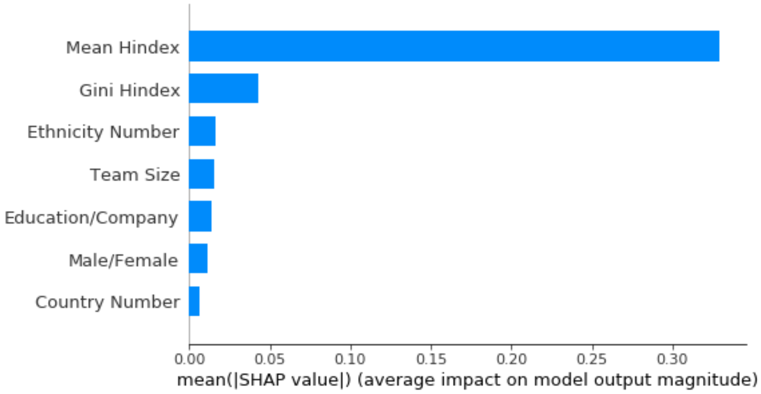

Figure 3. The average SHAP value magnitudes on DBLP dataset

Individual Feature Contribution

In Fig 4, each point represents the specific contribution value of each feature in each individual. The x-axis means the impact of each feature on the team's performance in citations, positive or negative. The color of each point implies the scale of the feature value (red is highest, blue is lowest). Firstly, as to the Mean H-index, the higher the mean H-index is, the more positive impact it has on team citation. When the Gini coefficient is low in H-index, which suggests that everyone shares a similar H-index, the team has a higher probability of having above-mean citations. When teams are more diverse in race, they have more citations. When the team size is larger, teams have more influence. Most of the combined teams (mixed color) and industry teams (pure red) are on the left side, whereas academia teams (pure blue) are on the right side. It shows that combined teams and industry teams outperform pure academia teams. Female-dominated teams (pure red), and equal teams (mixed color), which are mainly on the left, indicate that female-dominated teams and equal teams have fewer citations than male-dominated teams (pure blue). Finally, when teams are diverse in the country, they might have more citations or fewer citations. It suggests that country diversity can bring innovation but also bring cultural barriers.

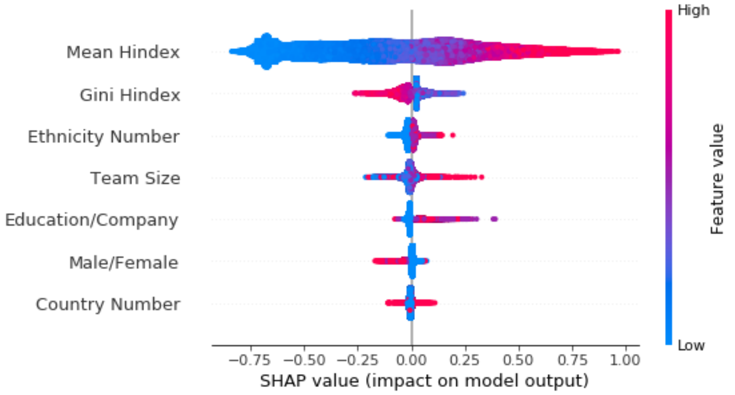

Figure 4. Individual feature contribution on DBLP dataset

Interactive Feature Contribution

We draw the dependence plots between any two variable pairs and found an interactive effect between team power hierarchy (Gini H-index) and team power level (Mean H-index) in Fig 5. Each dot is a team. The x-axis is the value of Gini H-index, whereas the y-axis is the SHAP value for the feature, which represents how much the feature's value changes the output of the model in prediction. The color bar represents a second feature Mean H-index. The blue points (low Mean H-index) are higher than red points (high Mean H-index) when the Gini H-index increases (Fig 5). When teams have a high hierarchy, teams with low team power levels are more likely to have a higher impact than teams with high team power levels (Fig 5). It suggests that teams with low team power level benefits more from a hierarchical structure than teams with high team power level.

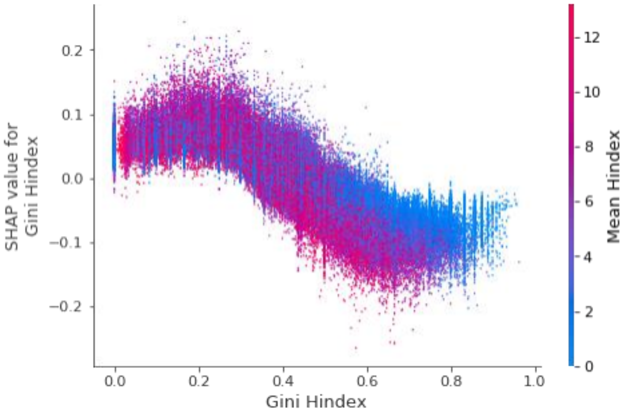

Figure 5. Interactive effects between Gini H-index and Mean H-index

CONCLUSION

This paper calculates different team composition and team structure variables to predict high/low team citation impact with the XGBoost model. Our model outperforms traditional models, such as the logistic model and random tree, in terms of ACC, F1, and AUC metrics. Then, we apply the Explainable AI SHAP method to interpret each feature contribution to each team's citation prediction. After aggregating all individual values, we found that team structure is more important than team composition. The team power level and hierarchy rank first and second in the overall contribution, followed by unique race number, team size, male/female dominance, academia/industry collaboration, and unique country number. High team power level is positively related to team citation, whereas high team power hierarchy is negatively related to team citation. There is an interactive effect between team power level and hierarchy. When team members are more diverse in race, are more in team scale, and have academia/industry interaction, they can receive more citations. But including female collaborators can reduce citations. It might be caused by the specific CS discipline where female researchers are underrepresented. The diversity in institutional countries cannot guarantee an increase in citations.

Our project makes important contributions to team citation prediction from team composition and structure. We enrich team-related variables rather than emphasize individual superstars to predict citation impact. The ranking of feature prediction through SHAP values can offer suggestions to scientific teams about the priority lists when they try to form new teams and build connections to maximize team performance. However, we need to admit that the team-related variables we choose have limitations. On the one hand, variables about mean career age, mean productivity, mean citation and mean collaborations and mean H-index have an overlap to some extent. On the other hand, the variables we choose are not complete. We could still consider the team composition from the department dimension. Or we could consider whether teams belong to high-rank institutions. Secondly, our model is trained to predict whether team citation is above the mean value or not. The best model, XGBoost, can only reach 70% accuracy. The prediction results still have gaps with previous literature. Finally, we only test our results in the CS field. For the future work, there are interesting areas of extending this research to different disciplines including science, social science, and humanities. It is great to use two Explainable AI methods to cross validate the feature importance.

ACKNOWLEDGMENTS

This work was supported by NSF-CSIRO (NSF 2303038), NIH Bridge2AI (OTA-21-008), and Bill & Lewis Suit Endowment Fund at the University of Texas at Austin.

REFERENCES

Acuna, D. E., & Liang, L. (2021). Are AI ethics conferences different and more diverse compared to traditional computer science conferences?.

Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. The Quarterly Journal of Economics, 107(4), 1137-1160.

Berdahl, J. L., & Anderson, C. (2005). Men, women, and leadership centralization in groups over time. Group Dynamics: Theory, Research, and Practice, 9(1), 45.

Blau, P. M. (1964). Exchange and power in social life. New Brunswick.

documentation (2008).

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cole, S., & Cole, J. R. (1967). Scientific output and recognition: A study in the operation of the reward system in science. American sociological review, 377-390.

Freeman, R. B., & Huang, W. (2015). Collaborating with people like me: Ethnic coauthorship within the United States. Journal of Labor Economics, 33(S1), S289-S318.

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics, 85(1), 257-270.

Greer, L. L., & Van Kleef, G. A. (2010). Equality versus differentiation: The effects of power dispersion on group interaction. Journal of Applied Psychology, 95(6), 1032 1 044.

Greer, L. L., Caruso, H. M., & Jehn, K. A. (2011). The bigger they are, the harder they fall: Linking team power, team conflict, and performance. Organizational Behavior and Human Decision Processes, 116(1), 116 1 28

Greer, L. L., De Jong, B. A., Schouten, M. E., & Dannals, J. (2018). Why and when hierarchy impacts team effectiveness: A meta-analytic integration. Journal of Applied Psychology, 103, 591 6 13.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. Science, 308(5722), 697-702.

Halevy, N., Chou, E. Y., Galinsky, A. D., & Murnighan, J. K. (2012). When hierarchy wins: Evidence from the national basketball association. Social Psychological and Personality Science, 3(4), 398 4 06.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences, 102(46), 16569-16572.

Hofman, Jake M., et al. "Integrating explanation and prediction in computational social science." Nature 595.7866 (2021): 181-188.

Hofstra, B., Kulkarni, V. V., Galvez, S. M. N., He, B., Jurafsky, D., & McFarland, D. A. (2020). The Diversity I nnovation Paradox in Science. Proceedings of the National Academy of Sciences, 117(17), 9284-9291.

Hsiehchen, D., Espinoza, M., & Hsieh, A. (2015). Multinational teams and diseconomies of scale in collaborative research. Science advances, 1(8), e1500211.

Hu, Y. H., Tai, C. T., Liu, K. E., & Cai, C. F. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity. Journal of Informetrics, 14(1), 101004.

Jiang, S., Koch, B., & Sun, Y. (2021, April). HINTS: Citation Time Series Prediction for New Publications via Dynamic Heterogeneous Information Network Embedding. In Proceedings of the Web Conference 2021 (pp. 3158-3167).

Larivière, V., Gingras, Y., & Sugimoto, C.R. (2014). Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology, 66(7), 1323-1332.

Larivière, V., Macaluso, B., Mongeon, P., Siler, K., & Sugimoto, C. R. (2018). Vanishing industries and the rising monopoly of universities in published research. PloS one, 13(8), e0202120.

Lebeau, L. M., Laframboise, M. C., Larivière, V., & Gingras, Y. (2008). The effect of university i ndustry collaboration on the scientific impact of publications: the Canadian case, 1980 2 005. Research Evaluation, 17(3), 227-232.

Li, S., Zhao, W. X., Yin, E. J., & Wen, J. R. (2019). A neural citation count prediction model based on peer review text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 4914-4924).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. Scientometrics, 126(8), 6803-6823.

Ma, Yaxue, et al. "Identifying widely disseminated scientific papers on social media." Information Processing & Management 59.3 (2022): 102945.

Manjunath, A., Li, H., Song, S., Zhang, Z., & Kumar, I. (2021). Comprehensive analysis of 2.4 million patent-to-research citations maps the biomedical innovation and translation landscape. Nature Biotechnology, 39(6), 678-683.

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. University of Chicago press.

Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. isis, 79(4), 606-623.

Newman, M. E. (2014). Prediction of highly cited papers. EPL (Europhysics Letters), 105(2), 28002.

Roth, Camille, Jiang Wu, and Sergi Lozano. "Assessing impact and quality from local dynamics of citation networks." Journal of Informetrics 6.1 (2012): 111-120.

Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. Journal of Informetrics, 14(3), 101039.

Saveski, M., Awad, E., Rahwan, I., & Cebrian, M. (2021). Algorithmic and human prediction of success in human collaboration from visual features. Scientific Reports, 11(1), 1-13.

Sugimoto, C. R. (2021). Scientific success by numbers. Nature, 593(7857), 30-31.

Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. R. (2022). How status of research papers affects the way they are read and cited. Research Policy, 51(4), 104484.

Wagner, C.S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. Research Policy, 34(10), 1608-1618

Wagner, Caroline S., Travis A. Whetsell, and Satyam Mukherjee. "International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination." Research Policy 48.5 (2019): 1260-1270.

- 4 4.

Woolley, A. W., Gerbasi, M. E., Chabris, C. F., Kosslyn, S. M., & Hackman, J. R. (2008). Bringing in the experts: How team composition and collaborative planning jointly shape analytic effectiveness. Small Group Research, 39(3), 352-371.

Woolley, Anita Williams, et al. "Evidence for a collective intelligence factor in the performance of human groups." science 330.6004 (2010): 686-688.

Wu, Lingfei, Dashun Wang, and James A. Evans. "Large teams develop and small teams disrupt science and technology." Nature 566.7744 (2019): 378-382.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. "The increasing dominance of teams in production of knowledge." Science 316.5827 (2007): 1036-1039.

Xu, F., Wu, L., & Evans, J. (2022). Flat teams drive scientific innovation. Proceedings of the National Academy of Sciences, 119(23), e2200927119.

Xu, H., Bu, Y., Liu, M., Zhang, C., Sun, M., Zhang, Y., ... & Ding, Y. (2022). Team power dynamics and team impact: New perspectives on scientific collaboration using career age as a proxy for team power. Journal of the Association for Information Science and Technology.

Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011, October). Citation count prediction: learning to estimate future citations for literature. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1247-1252).

Yang, R. (2020). Who dies from COVID-19? Post-hoc explanations of mortality prediction models using coalitional game theory, surrogate trees, and partial dependence plots. medRxiv.

Yu, T., Yu, G., Li, P. Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. Scientometrics, 101(2), 1233-1252.