An Exploratory Study on Kaggle Competition Trend: Topics, Submissions, and User Performance

Chad Milburn [*] Bhanuprakash Narayana[†] Trevor Speich [‡] Namith Manjunath Telkar[§]

Haijing Tu [¶]
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington

Abstract

The Meta Kaggle dataset includes a set of tables that offer an excellent overview of the variety of data analytics, modelling, and visualization challenges in the age of big data [11]. This report utilizes Kaggle competition data from 2010 to 2024 to explore the trend of data competitions, including competition topics, programming languages, submissions, evaluation metrics, and factors associated with team performance.

Index Terms: Data Visualization, Kaggle Competition, Data Science Trend, Machine Learning, Evaluation Metrics, Competition Rewards, User performance

1 Introduction

Kaggle is the most well-known platform for data scientists and machine learning practitioners to compete and connect with each other. Since 2010, millions of competitors have participated in thousands of Kaggle competitions, which allows the platform to gather a large volume of data on its users, competitions, scores, and kernels in two separate datasets: Meta Kaggle [11] and Metal Kaggle Code dataset [10].

This paper aims to extract insights from the Meta Kaggle dataset [11] to visualize the evolution of AI and machine learning competitions from 2011-2024, including trending competition topics, the growth of machine learning libraries usage, the evolvement of evaluation metrics, and factors associated with team performance in the competitions.

2 Kaggle Competitions and Their Contributions

2.1 Kaggle Competitions Cases

Kaggle competitions typically involve training machine learning algorithms to build models and visualizations for better prediction and better understanding of the problems under investigation. These competitions are driven by refining algorithms, improving accuracy in image recognition and classification, and contribution to academic research.

2.1.1 Image Recognition and Image Classification

Image recognition is one of the most frequent topics for Kaggle competitions. In 2016, a Kaggle competition on the topic of ultrasound nerve segmentation challenged Kagglers to identify nerve structures in a dataset of ultrasound images of the neck [7]. There were 924 submissions to this competition, and the winning team yielded a dice coefficient score as high as 0.73. In 2020 and 2022, The HuBMAP MC-IU team hosted multiple competitions on segmentation tasks and received over 1000 submissions evaluated by Dice Coefficient scores[2, 3].

Other examples for image recognition include a TSA threat recognition competition in 2017 with 149 submission ranked by `log loss` [6] and a single-cell classification Competition in 2021 that had 758 submissions evaluated by `IoU` (intersection over union) [9].

2.1.2 Regression Models

Kaggler competitions also aim to build predictive models using regression algorithms. In 2017, a Kaggle competition on housing values predictions was launched for Zillow [1]. The competition aims at refining their Zestimate algorithm—a home value estimation model. The algorithm leveraged hundreds of data points for each property to predict accurate home values. It continuously reduced the margin of error in predictions, from an initial 14% down to 5% during the competition. In total, 3772 teams joined this competition, and 3 winners led the score board with a `log error` score of 0.074.

2.1.3 Kaggle Competitions and Their Academic Contributions

The impact of Kaggle competitions has extended far beyond data science practitioners and is now evident in academic publications. Top performers in Kaggle competitions are rewarded not only with money, but also publishing opportunities in academic journals such as *Nature Methods* [12]. For example, several academic papers [5, 4] are published on the Kaggle competition "Hacking the Human Body," which had 1175 teams worldwide [3].

3 Kaggle Competition Data Summary and Visualization

The Meta Kaggle Dataset [11] contains comprehensive data on Kaggle competitions data with 32 tables and a total size of 30.19GB. Based on the goals of this research, the following selected tables are used for data analysis and visualization(Tab. 1).

3.1 Descriptive Visualizations

An overview of the top competition themes, programming language and packages, evaluation metrics, and submission trends are provided in the descriptive visualizations in this section.

3.1.1 Competition Themes

Fig. 1 illustrates the diverse range of topics of Kaggle competitions. On the left, we see the most popular dataset names which include terms like 'earth', 'science', 'data', and 'health', signifying the critical sectors where data science is applied. Similarly, 'business', 'trained', and 'model' suggest a focus on practical, industry-ready

[*] e-mail: cwmilbur@iu.edu
[†] e-mail: bhnaraya@iu.edu
[‡] e-mail: tspeich@iu.edu
[§] e-mail: ntelkar@iu.edu
[¶] e-mail: haijing.tu@indstate.edu

Table 1: Selected Tables from Kaggle Dataset

Tables	Records	Attributes
Competitions.csv	5,662	42
CompetitionTags.csv	950	3
Datasets.csv	310,500	14
DatasetTags.csv	299,311	3
DatasetVersions.csv	1.05 million	14
ForumTopics.csv	383,590	3
KernelTags.csv	6.56 million	3
Kernels.csv	54.9 million	16
KernelLanguages.csv	10	4
KernelVersions.csv	173 million	19
Submissions.csv	13.8 million	10
Teams.csv	7.07 million	13
TeamMemberships.csv	12.3 million	4
Tags.csv	817	9
Users.csv	17.4 million	4

applications. On the right, the prominent competition names resonate with these themes, emphasizing areas like 'binary', 'multiclass', and 'tabular', which indicate the variety of machine learning tasks tackled by competitors.

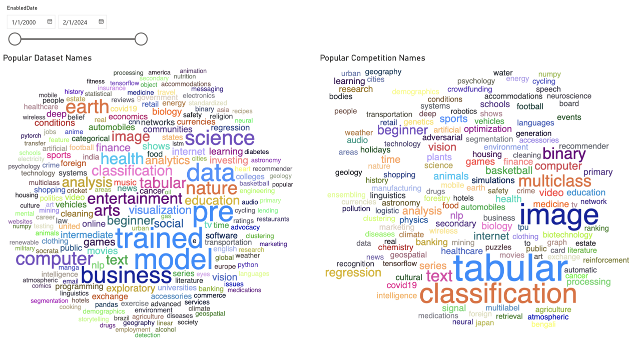


Figure 1: Word Cloud of Competition Topics(Interactive PowerBI Report)

Fig. 2 is a ribbon chart showing a trend of competition tags over the years and reflects the change of tag frequencies on the most popular tags such as earth and nature, health and fitness, and people and society, etc. Header tags were extracted for each tag to provide a clear and concise description of the tag categories.

3.1.2 Kaggle Trend in Programming Languages and Libraries

Fig. 3 utilizes a 100% stacked area chart to visualize the temporal evolution of script language usage. Each layer in the chart represents a different script language, showcasing their respective contributions over time. Python has been the dominating language since 2018.

Next, the use of Python libraries over the years is illustrated in Fig. 4. Python's appeal as a coding language stems largely from the powerful capabilities its libraries offer. Fig. 4 settled on the top 12 packages used throughout the years. Clearly there is a trend in the popularity of `sklearn` and `TensorFlow` with a steady decline of `NumPy`.

3.1.3 Evaluation Metrics

Fig. 5 shows the trend of top evaluation metrics over years. AUC (Area under the curve), Categorization Accuracy, and Root Mean Squared Error are among the top evaluation metrics. After normalizing the data, the percentages of competition metrics are calculated based on the frequency of competition metrics used in each specific year. As found in the visualizaiton, Categorization Accuracy remains the most popular metric since 2012.

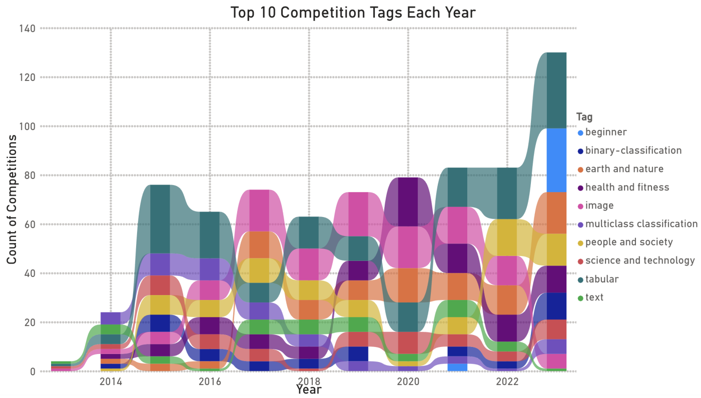


Figure 2: Ribbon Chart of Top 10 Common Competition Tags Each Year(Interactive PowerBI report)

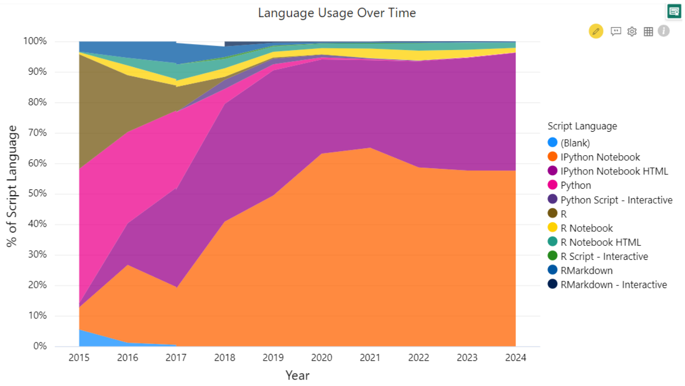


Figure 3: Trend Analysis of Programming Languages (Interactive PowerBI report)

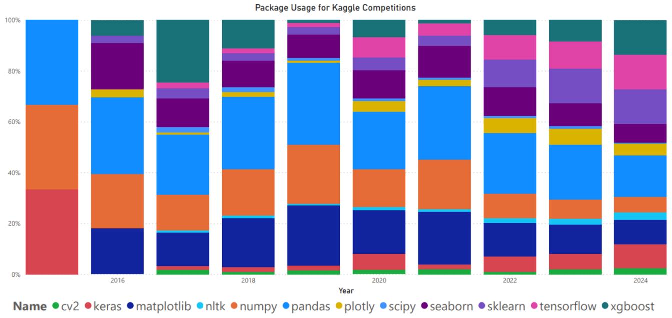
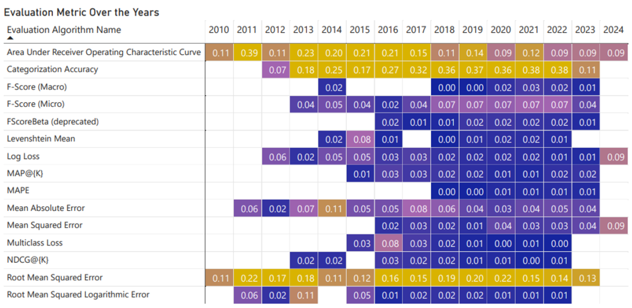


Figure 4: Trend Analysis of Python Packages and Libraries (Interactive PowerBI Report)

3.1.4 Competition Submission

Using a dual-axis bar chart, Fig. 6 provides an overview on the total number of Kagggle submissions and the average submissions per team over the top 10 competitions. Of the total 5,662 competitions included in the dataset, each competition had on average 132 teams,



Evaluation Metric Over the Years

Evaluation Algorithm Name	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Area Under Receiver Operating Characteristic Curve	0.11	0.39	0.11	0.23	0.20	0.21	0.21	0.15	0.11	0.14	0.09	0.12	0.09	0.09	0.09
Categorization Accuracy			0.07	0.18	0.25	0.17	0.27	0.32	0.36	0.37	0.36	0.38	0.38	0.11	
F-Score (Macro)					0.02				0.00	0.00	0.02	0.03	0.02	0.01	
F-Score (Micro)				0.04	0.05	0.04	0.02	0.04	0.07	0.07	0.07	0.07	0.07	0.04	
FScoreBeta (deprecated)							0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.01	
Levenshtein Mean					0.02	0.08	0.01		0.00	0.01	0.02	0.02	0.01		
Log Loss			0.06	0.02	0.05	0.05	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.09
MAP@{K}						0.01	0.03	0.03	0.02	0.02	0.01	0.01	0.02	0.02	
MAPE									0.00	0.00	0.01	0.01	0.01	0.01	
Mean Absolute Error		0.06	0.02	0.07	0.11	0.05	0.05	0.08	0.06	0.04	0.03	0.04	0.05	0.04	
Mean Squared Error							0.02	0.03	0.02	0.01	0.04	0.03	0.03	0.04	0.09
Multiclass Loss						0.03	0.08	0.03	0.02	0.01	0.00	0.01	0.00		
NDCG@{K}				0.02	0.02		0.03	0.02	0.02	0.01	0.01	0.00	0.01		
Root Mean Squared Error	0.11	0.22	0.17	0.18	0.11	0.12	0.16	0.15	0.19	0.20	0.22	0.15	0.14	0.13	
Root Mean Squared Logarithmic Error		0.06	0.02	0.11		0.05	0.01	0.02	0.02	0.02	0.01	0.01	0.01		

Figure 5: Evaluation Metrics over the Years (Interactive PowerBI Report)

and each team on average made 13.5 submissions.

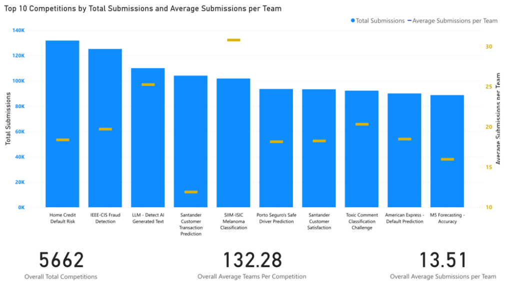


Figure 6: Submissions to Top Competitions and Submission Attempts Per Team(PowerBI Report)

3.2 Relationship Visualizations

Based on the results of descriptive summaries, this study further explores factors associated with team performance. These factors include numbers of submission, team sizes, participation in Kaggle forums, rewards types and reward amount.

3.2.1 Submission and Performance

Fig. 7 investigates the intricate relationship between the frequency of submissions and performance scores among the leading 100 teams across ten prominent Kaggle competitions. The plots divide the data into public and private leaderboards, reflecting the strategic adjustments teams make between initial submissions and final evaluations.

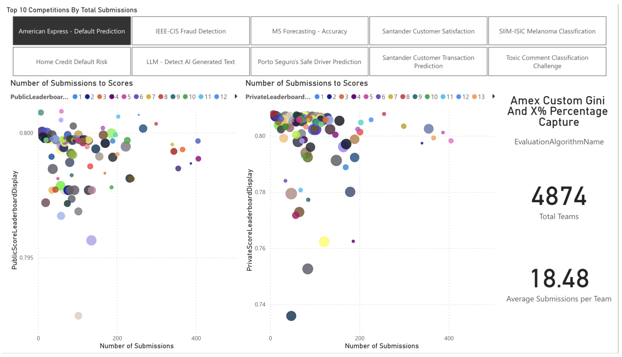

Top 10 Competitions By Total Submissions
American Express - Default Prediction
IEEE-CIS Fraud Detection
M5 Forecasting - Accuracy
Santander Customer Satisfaction
SIIM-ISIC Melanoma Classification
Home Credit Default Risk
LLM - Detect AI Generated Text
Porto Seguro's Safe Driver Prediction
Santander Customer Transaction Prediction
Toxic Comment Classification Challenge
Number of Submissions to Scores
PublicLeaderboar... ● 1 ● 2 ● 3 ● 4 ● 5 ● 6 ● 7 ● 8 ● 9 ● 10 ● 11 ● 12
PublicScoreLeaderboardDisplay
0.800
0.795
0
200
400
Number of Submissions
Number of Submissions to Scores
PrivateLeaderboard... ● 1 ● 2 ● 3 ● 4 ● 5 ● 6 ● 7 ● 8 ● 9 ● 10 ● 11 ● 12 ● 13
PrivateScoreLeaderboardDisplay
0.80
0.78
0.76
0.74
0
200
400
Number of Submissions
Amex Custom Gini And X% Percentage Capture
EvaluationAlgorithmName
4874
Total Teams
18.48
Average Submissions per Team

Figure 7: Top 10 Competitions and User Performance (Interactive PowerBI Report)

3.2.2 Team Size and Performance

Fig. 8 visualizes the relationship between team size and the number of competitions they participated. Single member teams are the most dominant over the years (98.4%), followed by two-member teams (0.96%), and teams with multiple members (0.64%). Therefore, we have two separate charts for single-members teams and multiple-member teams respectively.

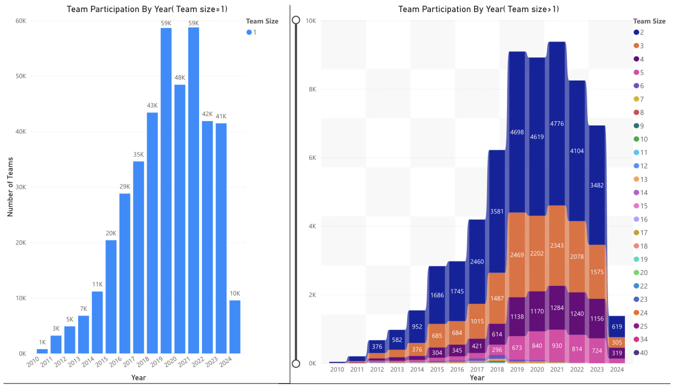


Figure 8: Team Size Distribution(Interactive PowerBI Report)

3.2.3 Forum Participation and Performance

Fig. 9 illustrates the number of ids in Kaggle Forums and their average votes from year to year, with the size representing the total number of the forum messages for each year. Recent years show significant growth in the number of forum messages.

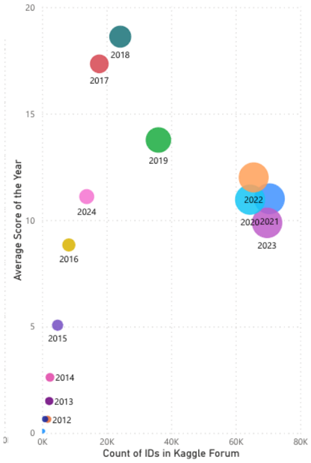


Figure 9: Forum Participation, User Score, and Sum of Total Messages by Year(Interactive PowerBI Report)

3.2.4 Competition Rewards and Performance

Fig. 10 shows a clear distinction between competitions that offer monetary reward versus those offering other rewards. Most submissions compete for monetary rewards, as numbers of submission clusters toward he lower end of monetary reward. Smaller monetary rewards seem to attract more submissions.

4 Discussion

4.1 Major Findings and Conclusion

There are promising signs indicating the tremendous growth in popularity and influence of Kaggle over its short 14-year existence. For

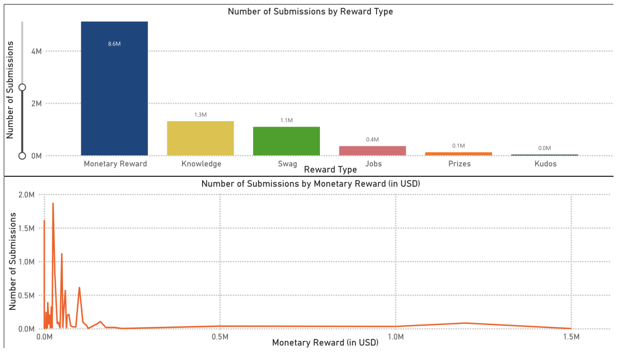


Figure 10: User Performance and Reward Type and Reward Amount (Interactive PowerBI Report)

example, the vast majority of Kagglers are beginners in their data competition journey (Tab. 2), At the same time, while the number of Kaggle users has been increasing, the number of IDs in the Kaggle forum has remained stagnant since 2019(Fig. 9). In the year of 2023, 4.5 million users joined Kaggle, an exponential growth compared with 4,509 registered users back in 2010. As a machine learning and data science competition platform, Kaggle records the trend of data analytics and visualization practices in which great shifts have been depicted through our report.

Table 2: Kaggle Users in Performance Tier

Performance Tier	Count of IDs
0	17,224,767
1	232,565
2	16,225
3	2,608
4	502
5	78

The visualizations developed in this paper well illustrated the switch from R to Python over the years and the fluctuation of popularity in Python packages.The rising popularity of `sklearn`, and `TensorFlow` indicate the wide use of machine learning, neural network, and deep learning in solving problems. In addition, certain tags consistently dominated competition themes in recent years, such as ”tabular” and ”image,” reflecting enduring trends in data science methodologies.

4.2 Challenges and Opportunities

It takes several major challenges to navigate through a big dataset like the Meta Kaggle Dataset. Building the model in PowerBI is difficult because these tables are not connected by primary and foreign keys, as ”id” columns in different tables have different labels. In addition, the many-to-many relationship between tables increases the size of data in PowerBI and challenges makes it difficult to process.

The visual complexity also arises from the differing scoring methodologies unique to each competition, where the same numerical score can signify contrasting levels of achievement. This underscores the competitive variability and the nuanced approaches teams must adopt to excel.

During validation, one of the problems surfaced was the lack of discernible and meaningful relationship between competition experience and users’ performance tier. It is also unclear what are the most predictive factors that contribute to competitors’ performance.

4.3 Future Study

People are more immersed in the world of data every day. In addition to the massive volume of data, there are many other challenges Big Data Analytics (BDA) poses for machine learning and data analysis, including ”format variation of the raw data, fast-moving streaming data, trustworthiness of the data analysis, highly distributed input sources, noisy and poor quality data, high dimensionality, scalability of algorithms, imbalanced input data, unsupervised and un-categorized data, limited supervised/labeled data, etc.” ([8], p.2). This study focuses more on the descriptive trends of Kaggle competition. Future studies should focus on building models predicting competitor’s performance, which will help competitors to find out their chance to solve a problem as they embark on the journey to make meaningful predictions using big data.

Supplemental Materials

Supplemental materials include PowerBi reports containing interactive data visualizations for all of the graphs included in this paper are accessible through the following links by requesting permission:

Competitions, Submissions, and Scores
Packages and Metrics
Languages, Top Competitions, Rewards and Performance
Team Size and Performance
Forum Participation and Performance

Acknowledgments

The authors wish to thank Dr. Andreas Bueckle and Professor Michael Ginda at Indiana University Bloomington for their guidance on this project. The authors also thank Mr. Yashvardhan Jain for his sponsorship and pivotal support, which kick-started our research.

References

[1] AndrewMartin, Bin, C. N, K. Nielsen, Maggie, and W. Kan. Zillow prize: Zillow’s home value prediction (zestimate), 2017. 1

[2] A. Howard, A. Lawrence, B. Sims, E. Tinsley, J. Kazmierczak, K. Borner, L. Godwin, M. Novaes, P. Culliton, R. Holland, R. Watson, and Y. Ju. Hubmap - hacking the kidney, 2020. 1

[3] A. Howard, C. Lindskog, E. Lundberg, K. Borner, L. Godwin, Shriya, S. Dane, T. Le, and Y. Jain. Hubmap + hpa - hacking the human body, 2022. 1

[4] Y. Jain, L. L. Godwin, S. Joshi, S. Mandarapu, T. Le, C. Lindskog, E. Lundberg, and K. Börne. Segmenting functional tissue units across human organs using community-driven development of generalizable machine learning algorithms. *Nature Communications*, 14, 2023. 1

[5] Y. Jain, L. L. Godwin, Y. Ju, N. Sood, E. M. Quardokus, A. Bueckle, T. Longacre, A. Horning, Y. Lin, E. D. Esplin, J. W. Hickey, M. P. Snyder, N. H. Patterson, J. M. Spraggins, and K. Börner. Segmentation of human functional tissue units in support of a human reference atlas. *Communications Biology*, 6, 2023. 1

[6] B. Lewis, M. McDonald, W. Bill, and W. Cukierski. Passenger screening algorithm challenge, 2017. 1

[7] A. Montoya, Hasnin, kaggle446, shirzad, W. Cukierski, and yffud. Ultrasound nerve segmentation, 2016. 1

[8] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1–21, 2015. doi: DOI 10.1186/s40537-014-0007-7 4

[9] W. Ouyang, C. Winsnes, and M. Hjelmare. Analysis of the human protein atlas image classification competition. *Nat Methods*, 16:1254–1261, 2019. 1

[10] J. Plotts and M. Risdal. Meta kaggle code, 2023. doi: 10.34740/KAGGLE/DS/3240808 1

[11] M. Risdal and T. Bozsolik. Meta kaggle: Kaggle’s public data on competitions, users, submission scores, and kernels. Kaggle, 2022. doi: 10.34740/KAGGLE/DS/9 1

[12] C. Winsnes, D. Sullivan, E. Park, E. Lundberg, Maggie, M. Hjelmare, and P. Culliton. Human protein atlas image classification, 2018. 1