Understanding Kaggle Competitions: A Comprehensive Examination of Meta Kaggle Datasets

Niteesha Jangam
Indiana University

Rasaghna Kuturu
Indiana University

Yamini Ane
Indiana University

Geetika Elaprolu
Indiana University

Pranay Reddy Gundala
Indiana University

ABSTRACT

The project on analysis of the meta Kaggle data presents an in-depth analysis of the transformative trends within Kaggle's Machine Learning and Data Science competitions over the past decade. Through meticulously crafted infographics, this project explores the evolution of competition participation, the dynamics of team collaborations, and the influence of prize incentives on competitor engagement by delving into the progression of solution methodologies, emphasizing the critical role of popular libraries and packages. Additionally, the examination of leaderboard fluctuations offers insights into the competitive landscape's inherent volatility. By distilling complex data into a coherent narrative, this study aims to elucidate the principal factors that have steered Kaggle's journey, fostering a culture of innovation and growth in the data science arena.

INTRODUCTION

In recent years, Kaggle has emerged as a very prominent platform for data science and machine learning enthusiasts to collaborate on projects, compete in competitions, and have access to a vast variety of datasets. This research focuses on finding insights extracted from the examination of the Meta Kaggle datasets. The areas of focus include participation trends, the impact of reward incentives, the evolution of team dynamics, the analysis of forum activity, and the efficiency of evaluation algorithms. Tableau and Python are the main tools that have been used for this study.

1.1 Description of the data

The Meta Kaggle data contains several datasets. The main focus of these datasets is on competitions, organizations, kernels, forum activity, and datasets. Table 1 shows a clear statistical summary of all the datasets.

Dataset	**Rows**	**Columns**
Competitions	5662	42
Kernels	1127909	16
Datasets	312769	14
Forums	339690	3
ForumMessages	4472212	18
ForumTopics	391550	13
Submissions	1048575	11
Organizations	3984	5
Teams	1048575	14

Table 1: No. of rows and columns in significant datasets

There are also connecting tags that help establish relationships between various datasets.

LITERATURE SURVEY

A considerable body of research has explored Meta Kaggle data, offering valuable insights into competition dynamics. Existing visualizations, such as kernel density estimate (KDE) plots [1], bar charts [3], and scatter plots (4), illuminate various aspects of Kaggle competitions. For instance, KDE plots [1] reveal score distributions and potential thresholds affecting rankings, while bar charts illustrate competition popularity and community interest in different problem domains. Scatter plots, possibly derived from Principal Component Analysis (PCA), highlight variability and trends in competition data [4]. Notably, prior work often leverages interactive platforms like Tableau and Python libraries like Matplotlib, Seaborn, and Plotly for dynamic visualization. These approaches emphasize the importance of selecting relevant datasets, employing advanced visualizations, and rigorous pre-processing to uncover meaningful insights [2]. By building upon these methodologies, this study aims to deepen understanding and reveal new insights into Kaggle competitions, enhancing the field of data science.

DATA ANALYSIS

The data analysis part involves three sub-components; data processing, data modeling, and data visualization.

3.1 Data Processing

In the data cleaning and preprocessing phase, several steps were undertaken to ensure the integrity and usability of the datasets. Initially, duplicate records were identified and systematically removed based on key attributes to enhance data quality and maintain uniqueness. Memory optimization techniques were employed to handle large datasets efficiently, including the removal of unnecessary columns and those with significant missing data percentages. Additionally, data types were adjusted to more memory-efficient formats, such as converting numeric types to smaller ones where applicable. Boolean columns were transformed from TRUE/FALSE to 1/0 to reduce memory consumption and facilitate numerical analysis. Missing data below 5% was imputed using median imputation to preserve data integrity. These processes were conducted using Python's Pandas library for data manipulation and NumPy for numerical operations, ensuring thorough and effective data preprocessing.

3.2 Data Modeling

For data modeling in this project, various datasets are linked using tag IDs, facilitating comprehensive analysis across different aspects of Kaggle competitions. Tag IDs serve as common identifiers, enabling the integration of information from competitions, datasets, forums, submissions, and organizations. This interconnected approach allows for the examination of relationships between competition themes, dataset usage, forum discussions, user engagement, and organizing entities. By leveraging tag IDs as linking mechanisms, the data modeling process ensures a cohesive understanding of Kaggle's ecosystem, facilitating insightful analysis and decision-making.

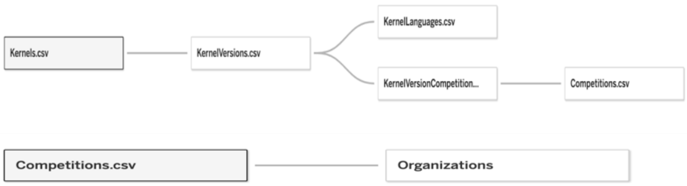

Figure 1. Various relationships are established to facilitate insightful analysis.

3.3 Data Visualization

For the data visualization section, we employed Tableau, PowerBI, and Python to address key research questions. Through interactive dashboards and insightful plots, we explored various aspects of Kaggle competitions. Questions tackled include changes in participation efficiency over the past decade, the impact of rewards on participation, and variations in reward offerings across competition types. We also investigated the evolution of evaluation algorithms and their growth across different competition types. Additionally, we analyzed the popularity of programming languages, sentiment perception in online communities, and variations in dataset engagement over different years, providing valuable insights into Kaggle's dynamics and trends.

RESULTS

4.1 Participation Efficiency

The data analysis and visualization of the Meta Kaggle data delve into the participation efficiency trend of competitions on the Kaggle platform over the past decade.

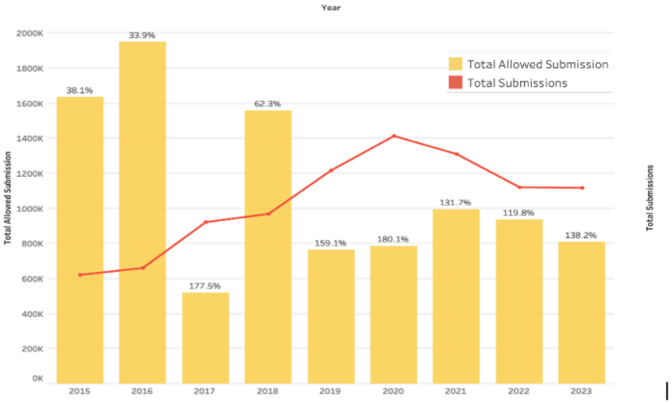

Figure: Participation efficiency of competitions over the past decade

Figure 2. Temporal analysis of the relationship between total submissions and allowable submissions

Figure 2 shows the relationship between total submissions and allowable submissions, offering insights into the effectiveness of competition rules and regulations in fostering active participation.

$Participation\ Efficiency\ (\%) = (Total\ Submissions / Total\ Allowed\ Submissions) * 100$

Total Allowable Submissions = $Total\ Competitors\ * Max\ daily\ Submissions$

Participation efficiency, calculated as the ratio of total submissions to total allowable submissions, reveals trends indicative of community engagement and competition dynamics. The observed downward trend in participation efficiency since 2019 prompts considerations of various factors such as competition format changes, the emergence of competing platforms, and shifting participant priorities.

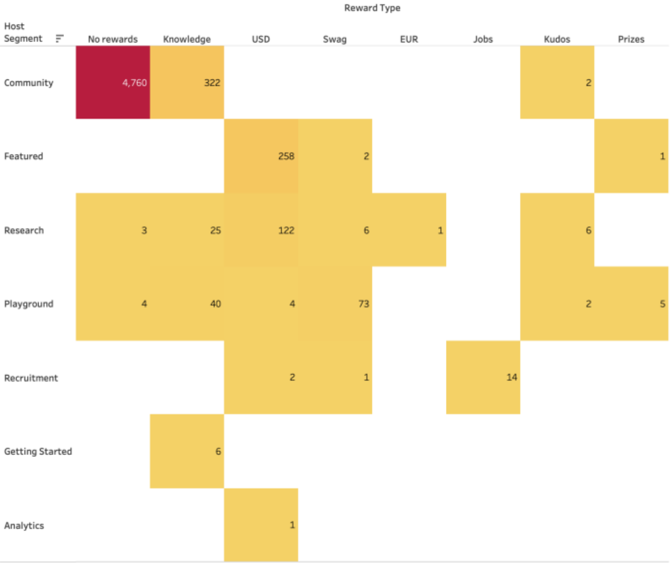

Figure 3. Change in participation according to the competition segment and rewards offered.

Competitions are held in organizations across various segments like research, recruitment, playground, getting started, community-based competitions, analytical and featured competitions. Participation in competitions is also determined by the segment in which the competitions are being conducted as participants also tend to choose competitions according to their interests and the end goal of the competition. The crosstab in Figure 3 shows how the reward type across different competition segments affects the number of competitors in the competition. Out of the 5662 competitors' data that are available, the highest number of competitors are observed in the community competitions having no rewards or knowledge as the outcome. This is followed by the 258 competitors in featured competitions where the rewards were in US dollars. Most of the research-based competitions having cash rewards also have a good number of competitors. One noticeable observation is that beginner-level 'Getting Started' competitions have only knowledge as their reward type. This trend suggests that these competitions are primarily focused on fostering learning and skill development among participants. While recruitment-based competitions majorly offer jobs or swags as rewards, playground competitions offer all kinds of rewards and also have considerably consistent competitors across all reward types.

4.2 Team Dynamics

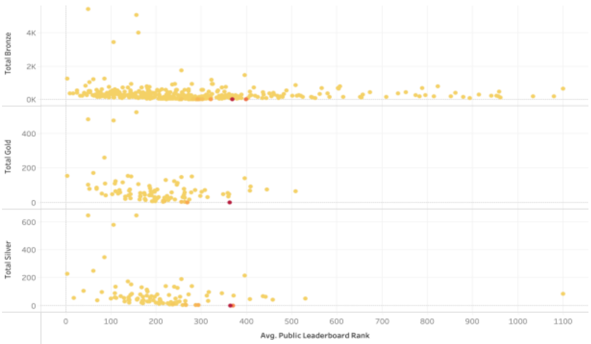

Figure 4. Change in the Average Public Leaderboard Ranking of Teams with the team member's achievement.

Team dynamics heavily depend on each member's performance. Figure 4's scatter plot shows that team members with many bronze medals usually have lower ranks, suggesting bronze medals are easier to get. Also, teams with lower average ranks win fewer medals. This means high-performing individuals significantly boost their team's success, demonstrating that a few top performers can greatly improve a team's results.

4.3 Change in Evaluation Algorithms

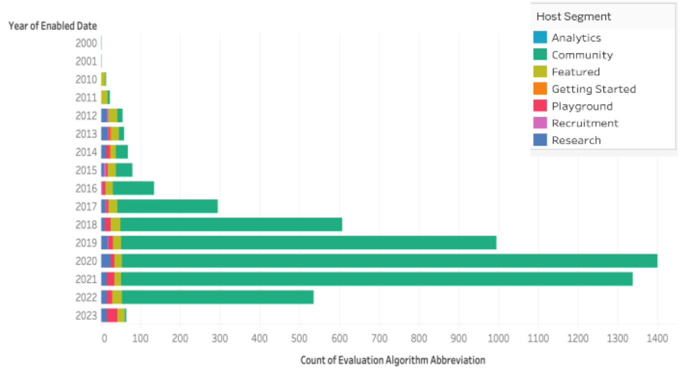

Figure 5. Usage of evaluation algorithms per year by host segment.

4.4 Programming Languages

The bar charts in Figure 6 answer the question about the popularity of programming languages on Kaggle by demonstrating that kernels using Python are generally more popular than R in terms of both engagement (votes) and interest (views), with this trend consistent across different types of competitions. The exception where R kernels slightly lead in views for Getting Started and Recruitment competitions might indicate specific contexts where R's usage is particularly relevant or sought after.

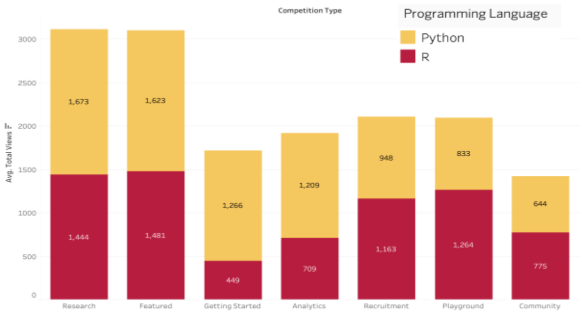

Figure 6. Average views of kernels across various competition types

4.5 Sentiment and Subjectivity in Forum Discussions

Examining the vocabulary used in Kaggle forum discussions reveals phrases related to collaboration and data science, indicating a supportive atmosphere within the community. Analysis of post sentiments shows that participants generally express positive feelings and contribute a mix of useful facts and personal experiences. This analysis clarifies the nature of engagement in the Kaggle community, demonstrating an environment rich in support and collaborative exchange.

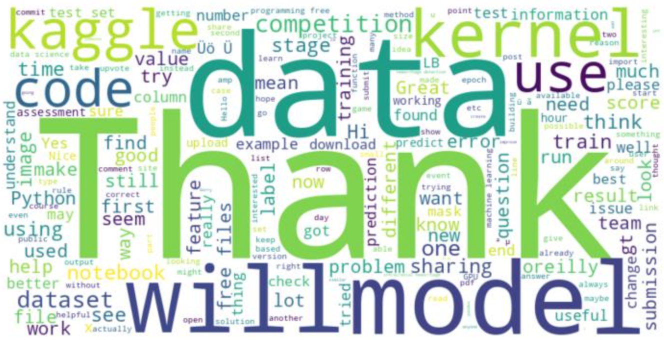

Figure 7. Word Cloud of Kaggle Forums

4.6 Dataset Engagement Trends

'Total Downloads' and 'Total Views' of Kaggle datasets from 2016 to 2023 are contrasted in the clustered column chart shown in Figure 8. Orange bars indicate views and yellow bars indicate downloads; the bars are presented on a vertical scale with counts in millions. 2020 saw a notable spike in views, followed by a fall in downloads and views after 2020. A larger inclination for viewing over downloading datasets is indicated by the consistently lower numbers of downloads relative to views, with the peak in 2020 suggesting a notable increase in user involvement during that time. It is also observed that there is a spike in the number of views of the datasets in the pandemic years of 2020 and 2021.

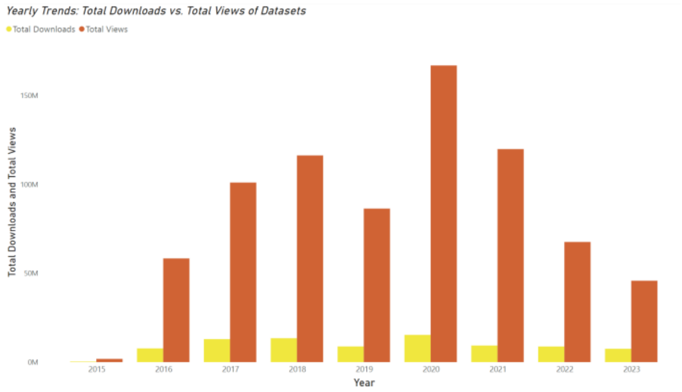

Figure 8. Temporal analysis of total downloads vs total views of datasets used in Kaggle.

Discussion

The analysis of Meta Kaggle data provides valuable insights into the trends and dynamics of Kaggle's Machine Learning and Data Science competitions. This study identifies key factors shaping Kaggle's evolution over the last decade by meticulously examining and visualizing competition participation, team collaborations, reward incentives, and evaluation algorithms. The analysis shows a downward trend in participation efficiency since 2019, prompting consideration of a variety of factors, including competition format changes, and shifting participant priorities. Further investigation reveals that participation levels vary across reward types, with cash rewards and knowledge-based competitions generating the most interest. Individual performance has a strong influence on team dynamics, with high-performing team members significantly contributing to their team's success. Furthermore, analysis of programming language preferences, top competition hosts, forum sentiment, and dataset engagement trends provides a thorough understanding of Kaggle's community dynamics and user engagement patterns. Overall, this study helps to shed light on the transformative trends in Kaggle's competitive landscape and promotes a better understanding of the factors that drive participant engagement and knowledge sharing within the data science community.

Conclusion

In conclusion, the analysis of Kaggle competitions indicates that teams led by strong leaders and proficient coders tend to achieve the best results. The need for specialized evaluation methods is increasingly critical as data science challenges evolve. Recommendations for future competitions include altering formats and diversifying reward structures to boost participation. Promoting teamwork and expertise in essential tools may also enhance competitor success. Moving forward, future research could explore the impact of emerging technologies, such as deep learning and reinforcement learning, on competition dynamics, as well as delve into the effects of external factors, such as global events, on participant engagement. Such investigations would further enrich our understanding of the data science community and its evolving landscape.

Acknowledgment

The authors wish to thank Kaggle and the larger online community for their free datasets and wisdom sharing, which were invaluable to our effort, and also appreciative of Yashwardhan Jain, and Andrews Buckle whose assistance was crucial to the accomplishment of this project.

References

[1] Tufte, Edward R. (1983). The Visual Display of Quantitative Information. Graphics Press.
[2] Moniruzzaman, M., Hossain, S. A., & Andersson, K. (2019). A survey of sentiment analysis techniques in English and Bangla text. Computer Science Review, 34, 100–124. https://doi.org/10.1016/j.cosrev.2019.02.001
[3] Kaggle. (n.d.). Meta Kaggle. Retrieved from https://www.kaggle.com/kaggle/meta-kaggle
[4] Nussbaumer Knaflic, Cole. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley.
[5] ScienceDirect. (n.d.). Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207019301189
[6] The Analytics Edge. (Spring 2015). Retrieved from https://www.kaggle.com/blobs/download/forum-message-attachment-files/2386/scores.png