Impact Score Technical Report

RMDS

Impact Score Team

Jiaxu (Carlo) Chen

Vincent Chen

Kooha Kwon

Shuyuan Wang

Peter (Zhipeng) Ye

Mo Chen

March 2021

Impact Score Technical Report

By RMDS

March 2021

1 Introduction

Impact score is a system that measures a data scientist's practical impact in their field. We believe that a data scientist's value begins with tangible impact through successful implementations of end-to-end projects and quick transition-to-market. Calculated using public records and platform user data, impact score demonstrates an individual's power or capacity to effect positive change in their field.

Some use cases of impact score can be:

- Technical recruiters - to source potential candidates on the platform and use impact score for initial screening
- Executives - to quantify their department's performance and talent pool with impact score
- Researchers - to select a team of engineers and analysts with a high impact score to support the implementation and commercialization of their research
- Academic institutions - to bridge the gap between academia and the application of industry knowledge by selecting professors who have high impact scores
- Data scientists - to seek out peers and mentors with high impact scores for project collaboration and review

One important purpose of this technical report is to present the reliability and validity of impact score, which are shown in detail through data analysis in the following sections.

2 Methods

2.1 Reliability Analysis

The investigation of impact score reliability focuses on the differences between the distributions of each version. Although impact score's algorithm is updated and improved from time to time, it is expected to be relatively consistent over the short term. In other words, we don't intend to update the algorithm in a way that causes the majority of the users to have a significant change in their impact scores between neighboring versions. Currently, there are two versions of impact score: v0.1 (developed in 2019) and v0.2 (developed in 2020). In this reliability analysis, we mainly focus on the difference between these two versions.

We start with the general statistics, especially the mean for each metric and for each group. These statistics gave us a sense of overall change in the scores. We then dive deeper into the dataset and locate

specific groups being influenced through a clustering analysis. After we identify the main types of changes, we dig deeper into the dataset and find out the reasons behind the change in the scores.

2.1.1 General Statistics

The main goal of general statistics method is to summarize the change. We first identified the common users of the old score and the new score (a couple of users churned while a few newly joined). Then, we calculate the change in each of the five categories defined in our impact score metrics: participation, capacity, project quality, project outcome, and project engagement.

$$\delta_{category} = new_{category} - old_{category}$$

We then calculated the statistics, including count, mean, standard deviation, and quartile, on them. There is no NaN value, but if there is, we drop that record from our table.

2.1.2 Clustering and Investigation

The clustering analysis finds the pattern within the score change table. We performed distribution analysis to cluster our customers. While a clustering algorithm is eligible for this case, our data shows a strong pattern, and we could easily identify all the change patterns easily through the distribution plot. We saved the step of clustering analysis and focused on investigating the root reason for the change. For each of the patterns detected, we joined the new score table and old score table and compared the specific sub-categories and their scores, and searched for changes exhaustively. After that, we summarized the root cause for each change pattern, specifically those that reduced the total score.

2.2 Validity Analysis

The main goal of the validity analysis is to check whether we can observe a correlation between a data scientist's impact score and their performance on other popular platforms. Such correlation, if it exists, is a good validation that impact score measures the real-world impact of data scientists.

In order to enact this comparison, we first needed to gather our users' profile information on outside platforms. In this analysis, user data is on four platforms: GitHub, LinkedIn, ResearchGate, and Kaggle. These four platforms are considered the most popular ones that data scientists visit. For users who opted to provide their profile URLs on these platforms, they are directly used. For those who did not provide their profile information, a web scrapping tool was developed (available at: https://GitHub.com/CarloCHEN/RMDS) to collect such data. The web scrapper is based on a Google API and searches the name and email registered at RMDS on each of the four platforms and return the found profile URLs. Due to people with same names or using a different email on these platforms, there is a chance of mismatch between the actual user and his/her profile URLs. The current matching rate is estimated to be 70%. Validation methods are being developed to enhance the matching rates and will be presented in the next version of the technical report.

We divided the validity analysis into two parts: the basic descriptive statistical analysis and the analysis with some filtering in the user data to avoid outlier issues. We first used all available data in each of the four websites respectively and conducted correlation analysis. Next, we used several filters to make our data more reliable. We first filtered the users with either 0 scores or 0 metrics in each of the websites, and conducted regression analysis. Then, we applied IQR method and eliminated the outliers in each of the metrics for each website and conducted regression analysis once again.

$$Q1 - 1.5 * IQR < x < Q3 + 1.5 * IQR$$

For GitHub, we are also aware that three types of users are in our group (use our platform often only, use GitHub often only, and use both often); therefore we clustered all metrics for GitHub and total impact score using 3 clusters (n_clusters = 3) and visualized them in scatterplot to validate that hypothesis.

3 Results and Discussion

3.1 Reliability

Our main observation in reliability analysis is that the difference between v0.1 and v0.2 for most users is small (on average decrease by 0.16). Impact score changes between these two versions are considered **stable and valid**.

3.1.1 Basic Statistics

On a very high level, our impact score has about 2,855 records for our old score system and 2,836 for our new score system. The average scores are both about 1.1 to 1.5.

	score	*quality*	*engagement*	*outcome*	*participation*	*capacity*
count	2855	2855	2855	2855	2855	2855
mean	1.41	0.09	0.07	0.01	0.24	1.00
std	2.96	0.72	0.65	0.15	1.08	1.81
min	0	0	0	0	0	0
25%	0.01	0	0	0	0.01	0
50%	0.03	0	0	0	0.01	0
75%	3.6	0	0	0	0.04	3.6
max	59.35	15.19	21.43	5.99	13.96	10.74

Fig 1: statistics for old score

	quality	*engagement*	*outcome*	*capacity*	*participation*	*score*
count	2836	2836	2836	2836	2836	2836
mean	0.13	0.04	0.05	0.69	0.25	1.16
std	1.01	0.36	0.48	1.23	0.69	2.53
min	0	0	0	0	0	0
25%	0	0	0	0	0	0
50%	0	0	0	0	0	0
75%	0	0	0	1.92	0	2.4
max	17.80	10.36	10.8	8.66	7.53	45.70

Fig 2: statistics for new score

Key findings:

1. Change in the user base: new score has discarded 48 users who have really low scores (Fig 3)

	quality	*engagement*	*outcome*	*capacity*	*participation*	*score*
mean	0	0	0	0.02	1.03	1.05

Fig 3: the mean of discarded users in old score

3.1.2 Overall Change Trend

The new impact score decreased overall compared to the old scores, particularly decreasing some people's scores (those who had a really low score) to 0. On average, the new score has been decreased by about 0.16.

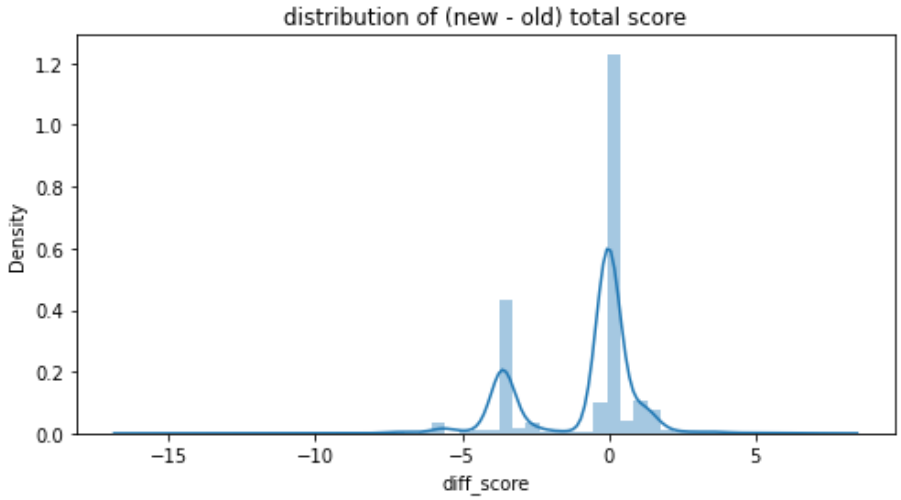

Fig 4: change of the score distribution

On average, the user experienced a slight increase in quality and outcome, but decreases in all other three dimensions.

	quality	*engagement*	*outcome*	*capacity*	*participation*
count	2807	2807	2807	2807	2807
mean	0.03	-0.04	0.04	-0.02	-0.87
std	0.30	0.39	0.55	1.00	1.61

Fig 5: group statistics for change in each dimension

3.1.3 Reduce Score Users Analysis

At a very general glance, the users who got reduced to 0 are those who originally had a low score.

One group of users who got reduced scores to 0 are those with only a few competitions and visitors but no other information. The new score might have an error in counting the competition and the visitors and might result in 0 in the new dataset.

Most users of this kind only had a 0.01 score originally in the old score, so the rounding function reduced them to zero automatically.

Another group of people got reduced in the metrics because some metrics are no longer available or changed definition. These features include num_click, which counts the total number of clicking in half a year. This metric has a different definition in the new score, so the user experienced a decrease. Another similar metric is the new_company or not. Since we changed the definition, they had a decrease.

3.2 Validity

The main observation in validity analysis is that there exists a **positive** correlation between impact score and ResearchGate performance, Kaggle performance, and some measurement of GitHub. No correlation is observed between impact score and LinkedIn measurement.

3.2.1 Correlation Between Impact Score and ResearchGate Measurement

There are 123 users with a ResearchGate account.

	uid	*quality*	*engagement*	*outcome*	*capacity*	*participation*	*score*	*Pub No.*
count	123	123	123	123	123	123	123	123
mean	2445.12	4.03	0.79	1.21	4.17	2.04	2.21	104.45
std	1170.24	11.87	2.81	6.63	10.09	4.05	4.53	217.53
min	333	0	0	0	0	0	0	2
25%	999	0	0	0	0	0	0	11.5
50%	2966	0	0	0	0	0	0	38
75%	3442	0	0	0	0	2.18	1.85	113
max	3812	45.78	17.56	58.56	57.35	21.54	26.22	1753

	Read No.	*Citation No.*	*avg_citation_per_pub*	*avg_read_per_pub*
count	123	123	123	123
mean	12778.26	2585.545	17.3548	157.1624
std	21753.15	7396.383	19.66835	201.1422
min	51	0	0	9.5
25%	1055	63.5	3.68	56.95
50%	4829	407	12.84	100.36
75%	13703	1900	23.015	185.035
max	127940	61879	112.14	1425.09

Fig 6: overall statistics for ResearchGate measurement and score

There are three measurements on ResearchGate, namely the number of the publication (Pub No.), the number of reads of the publication (Read No.), and the number of citations (Citation No.). Another two metrics are derived. Average read per publication (avg_read_per_pub) shows the exposure of the publication, and average citation per publication (avg_citation_per_pub) reflects the quality of the publication.

There exists positive relationships between score and Pub No., Read No., Citation No., and avg_read_per_pub, but negative relationship between score and avg_citation_per_pub. However, it is worth noticing that all these relationships are not statistically significant according to Pearson correlation test. Description and analysis here is only based on the absolute value of correlation coefficients.

	score	***Pub No.***	***Read No.***	***Citation No.***	***avg_read_per_pub***	***avg_citation_per_pub***
score	*1*	*0.060322*	*0.074456*	*0.053874*	*0.09319*	*-0.06296*
Pub No.	*0.060322*	*1*	*0.666833*	*0.933301*	*-0.0838*	*0.18212*
Read No.	*0.074456*	*0.666833*	*1*	*0.689081*	*0.438178*	*0.248406*
Citation No.	*0.053874*	*0.933301*	*0.689081*	*1*	*-0.05365*	*0.299646*
avg_read_per_pub	*0.09319*	*-0.0838*	*0.438178*	*-0.05365*	*1*	*0.125116*
avg_citation_per_pub	*-0.06296*	*0.18212*	*0.248406*	*0.299646*	*0.125116*	*1*

Fig 7: correlation between score and ResearchGate measurement

To dig deeper into this negative relationship, a scatter plot between score and avg_citation_per_pub is generated. On this plot, three outliers draw attention. Two with the highest and second highest avg_citation_per_pub have a 0 or nearly 0 score, and one with the highest score has a very low avg_citation_per_pub. These three points are considered as outliers because they are outside IQR.

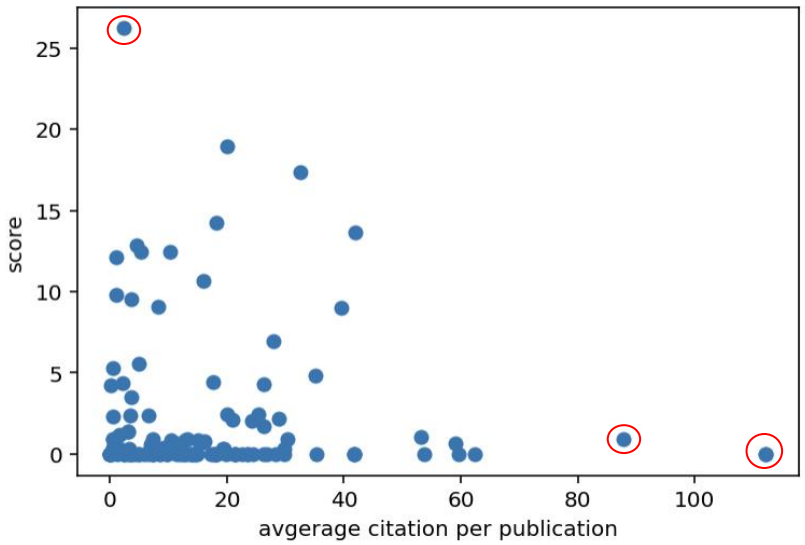

Fig 8: scatter plot of score versus average citation per publication

After the normalization of the number of citations (i.e., average the number of citations by the number of publications), the citation data is unified and therefore outliers may have a larger effect compared with the raw citation number. Ruling out the outliers, the relationship between score and avg_citation_per_pub becomes positive.

	score	***avg_citation_per_pub***
score	1	0.021617
avg_citation_per_pub	0.021617	1

Fig 9: correlation after eliminating outliers

Additionally, by observing the scatter plot, we can see that there are lots of points located near the x-axis, indicating a 0 or nearly 0 impact score. If we remove those points with 0 score, we see a positive trend.

Therefore, two insights can be drawn from the analysis between score and ResearchGate measurement. We need a larger sample size to get a more reliable result. We should encourage users to engage with the platform as much as we can so that there will be fewer 0 impact scores, and try new strategies to improve the user acquisition for the RMDS platform.

3.2.2 Correlation Between Impact Score and LinkedIn Measurement

There are 29 users with LinkedIn accounts. There is one numerical metric from LinkedIn measurement, i.e., number of connections; however, one should notice that the maximum number displayed on the LinkedIn network is 500. That is, we don't have access to the specific number of connections once the user has more than 500 connections.

	uid	***quality***	***engagement***	***outcome***	***capacity***	***participation***	***score***	***Number of Connections***
count	29	29	29	29	29	29	29	29
mean	1940.90	6.48	0.98	4.18	18.58	4.43	5.62	389.41
std	1045.48	14.77	2.50	11.28	12.11	7.31	6.15	154.36
min	539	0	0	0	0	0	0	94
25%	982	0	0	0	10	0	1.53	233
50%	1706	0	0	0	21.98	1.06	3.62	500
75%	2915	0	0	0	27.46	3.15	6.96	500
max	3515	49.21	9.92	50.64	46.62	25.73	21.68	500

Fig 10: overall statistics for LinkedIn measurement and score

There exists a negative relationship between number of connections and score. One important reason is that we don't have information for those with resourceful connections.

	uid	***quality***	***engagement***	***outcome***	***capacity***	***participation***	***score***	***Number of Connections***
uid	1	-0.30	-0.26	-0.23	-0.10	-0.46	-0.41	0.08
quality	-0.30	1	0.87	0.25	0.42	0.39	0.83	0.09
engagement	-0.26	0.87	1	0.40	0.30	0.32	0.78	-0.07

outcome	-0.23	0.25	0.40	1	0.26	0.18	0.60	-0.19
capacity	-0.09	0.42	0.30	0.26	1	0.27	0.63	0.16
participation	-0.46	0.39	0.32	0.18	0.27	1	0.66	-0.23
score	-0.41	0.83	0.78	0.60	0.63	0.66	1	-0.07
Number of Connections	0.08	0.09	-0.07	-0.19	0.16	-0.23	-0.07	1

Fig 11: correlation between scores and LinkedIn connections

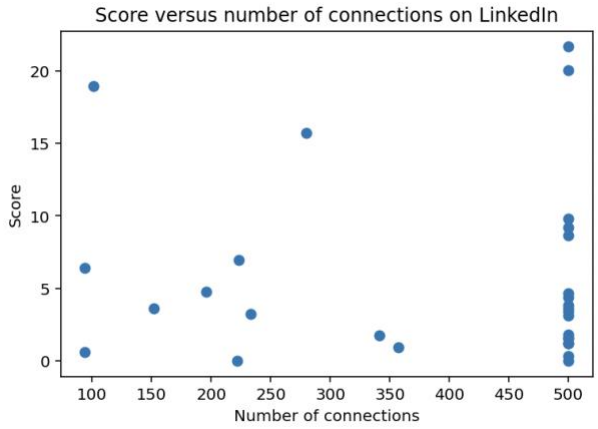

Fig 12: scatter plot for score versus LinkedIn connections

One possible way of validating is to obtain the number of followers under the Activity column in LinkedIn. If one is active on LinkedIn, the number of followers would generally be high, and this number doesn't have a maximal constraint as the number of connections.

3.2.3 Correlation Between Impact Score and GitHub Measurement

Currently, we have 94 samples for GitHub, and about 70 percent of them are reliable due to the naming issues. In other words, about 66 data points would be reliable and might incur skewness in our analysis. For example, for those accounts whose accuracy was uncertain, we found them to have little GitHub information, and may skew our statistics to the left.

By conducting correlation among users with the GitHub, we found the GitHub statistics have a weak positive in some measurements and a weak negative correlation in others with our current scores.

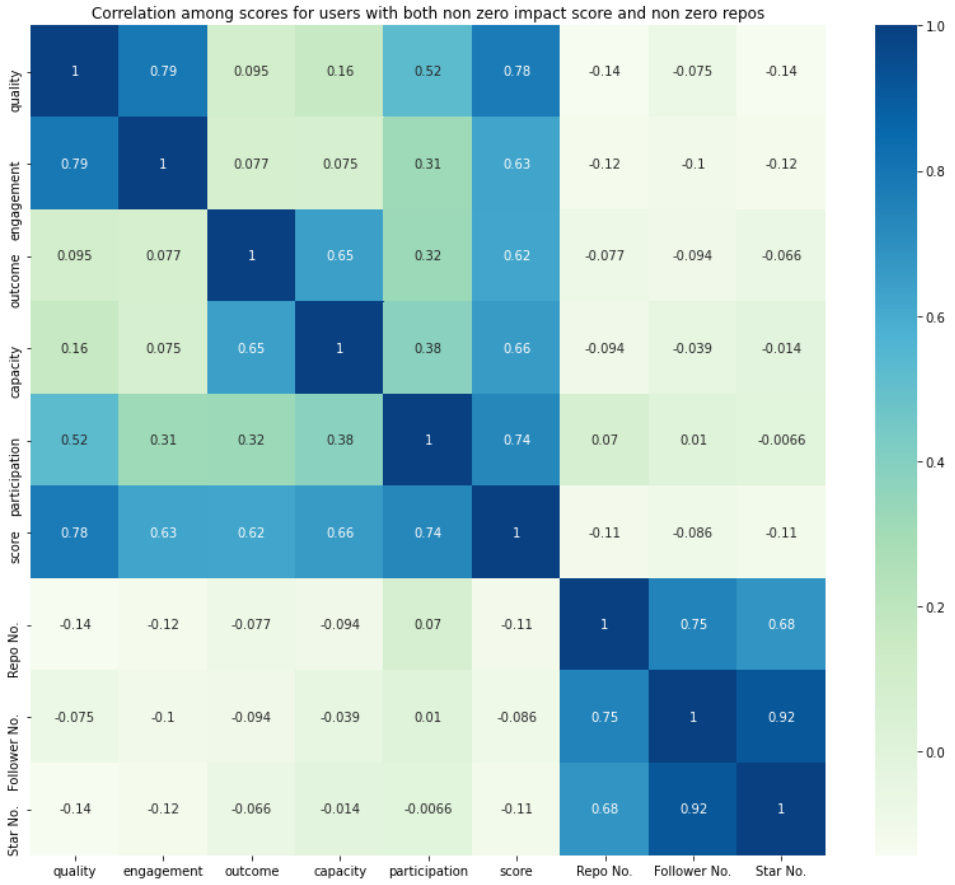

Fig 13: heatmap for correlation among GitHub and impact score features

Since we have a weak correlation, we plotted a scatterplot and found the impact of outliers (those with very high stars and low scores).

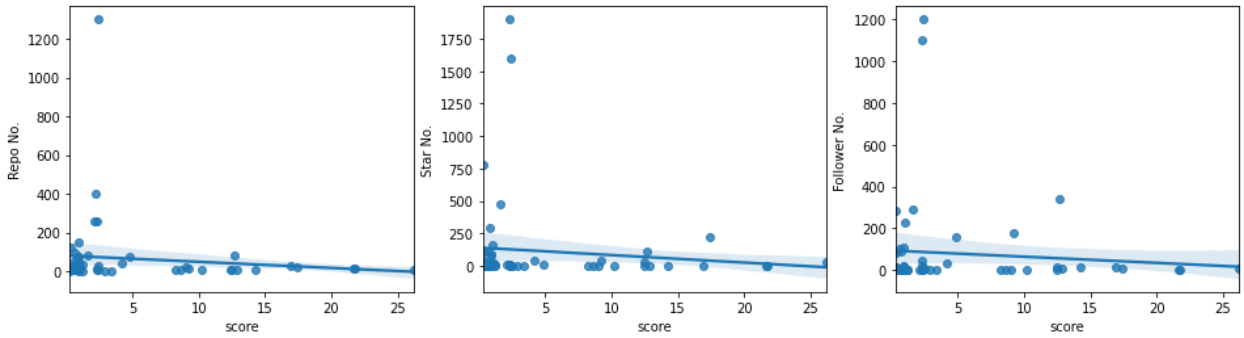

Fig 14: scatter plot between GitHub features and impact score before dropping outliers

The scatter plot indicates that we have a couple of outliers who might be skewing our analysis; therefore, we use IQR method to drop them (Q1 – 1.5 IQR < x < Q3 + 1.5 IQR). The new regression result is shown below.

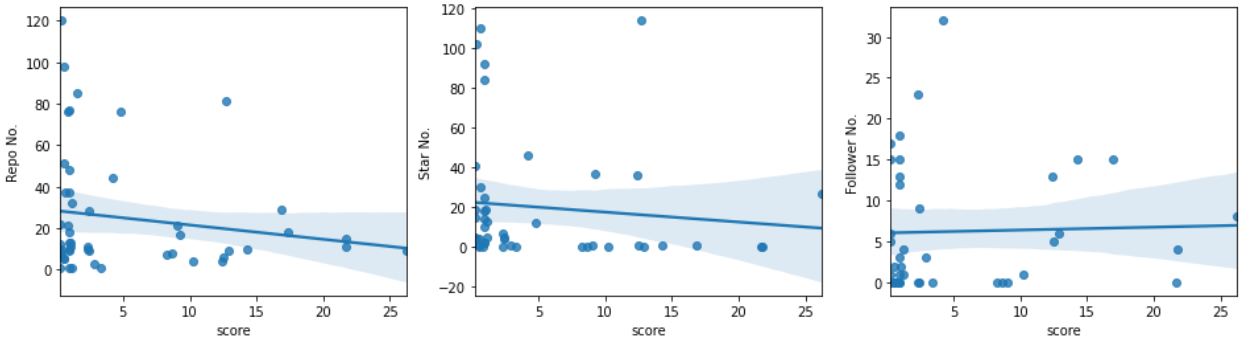

Fig 15: scatter plot between GitHub features and impact score after dropping outliers

However, there is an ambiguous direction in the correlation. This is hypothetically caused by multiple groups of people within the samples. For example, some are active in GitHub only, some are active in our platform only, and some are active in both; therefore, we tried K-Means on Repo, Star, Follower, and impact_scores metrics using 3 clusters and got the following results. The result seems to be aligned with our hypothesis. In addition to that, we found Star No. to be significantly negatively correlated with our scores.

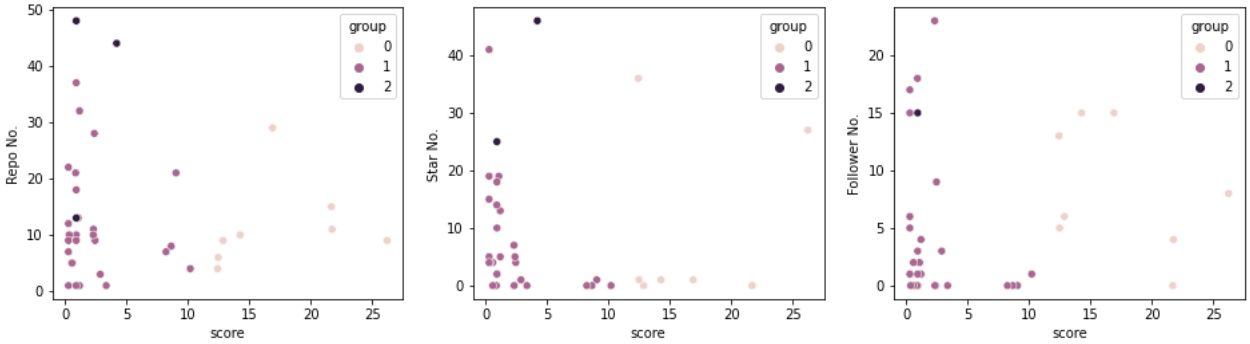

Fig 16: scatter plot between GitHub features and impact score with KMeans Clustering

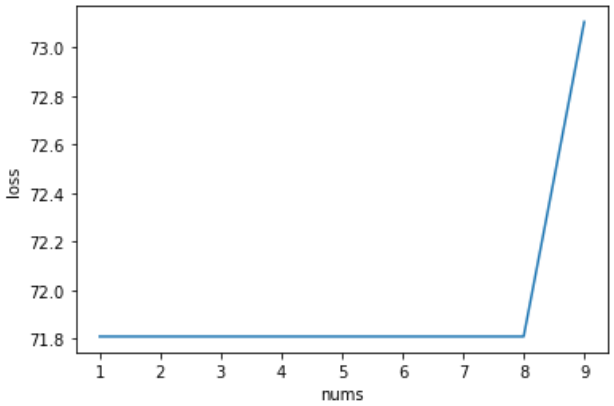

Fig 17: Elbow chart for K-Means Algorithm

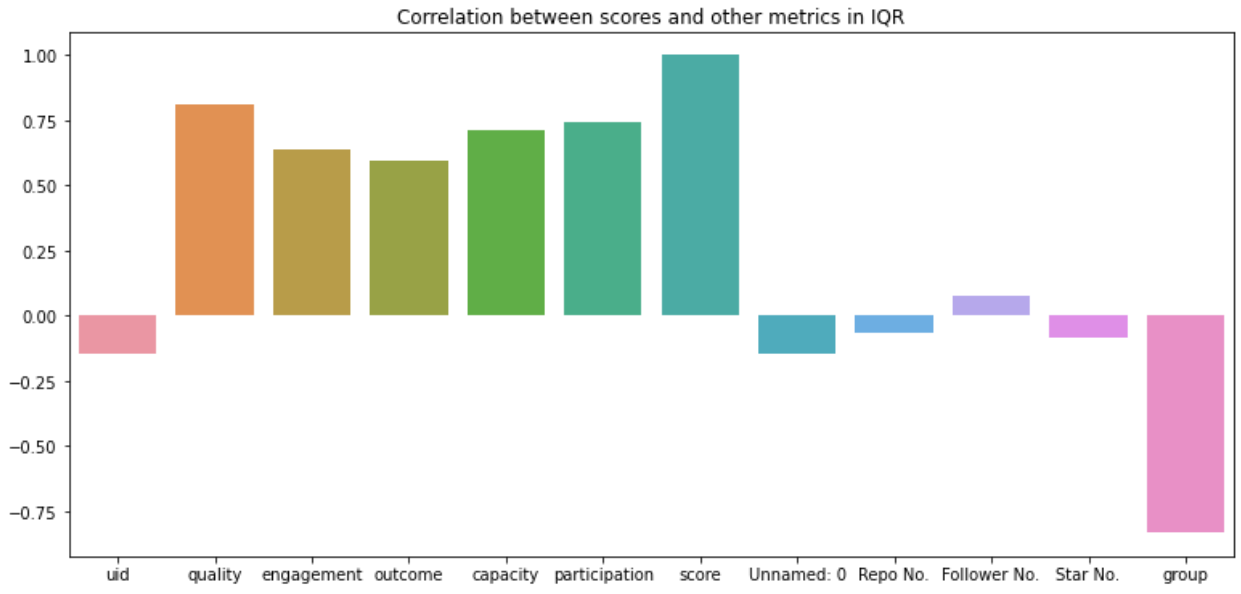

Fig 18: Correlation Bar chart between impact score and GitHub metrics after dropping outliers

3.2.4 Correlation Between Impact Score and Kaggle Measurement

Due to the current API limit, we only have three data points for Kaggle at the current stage. This issue is going to be fixed in the near future as we continue to scrape user information on a daily basis; however, the small sample size of this dataset might make our analysis non-robust and therefore we will not make many conclusions for now.

Currently, those who had Kaggle scores usually have a high impact score as well.

	uid	*quality*	*engagement*	*outcome*	*capacity*	*participation*	*score*
count	3	3	3	3	3	3	3
mean	422.33	29.57	0.79	16.68	21.43	10.72	15.18
std	327.42	25.65	6.45	14.45	28.89	9.18	9.57
min	110	0	0	0	0	0.13	9.10
25%	252	21.47	1.72	12.44	5	8.00	9.65
50%	394	42.94	3.45	24.87	10	15.87	10.20
75%	578.5	44.36	7.97	25.02	32.14	16.02	18.21
max	763	45.78	12.49	25.17	54.28	16.18	26.22

Fig 19: Basic statistics

On average, those 3 records have impact scores of 15 compared to 0.57 of those without Kaggle records; however, this might be skewed by the inactive account in the group without a Kaggle record.

While this may change in the future, we have a high correlation between impact score and competition No.

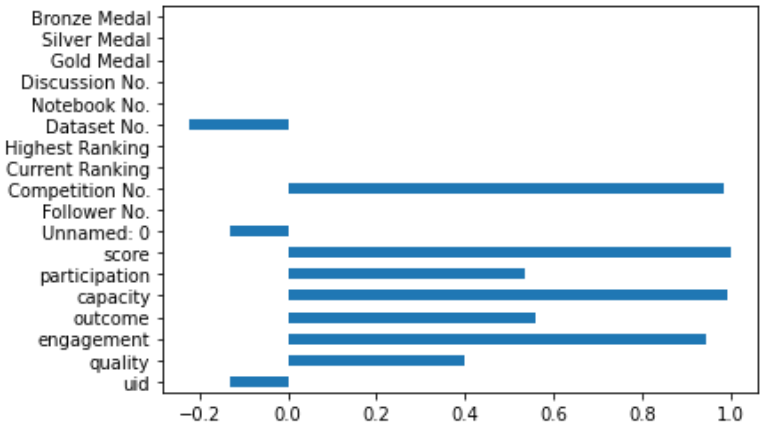

Fig 21: Correlation chart for impact score and Kaggle metrics

4 Conclusion and Future Expectation

The main observation in reliability analysis is that the difference between v0.1 and v0.2 for most users is small (on average decrease by 20%). Impact score changes between these two versions is considered stable, and the impact scores are reliable.

The main observation in validity analysis is that there are positive correlations between impact score and ResearchGate performance, Kaggle performance, and some measurements of GitHub, which all reflect the impacts of data scientists. No correlation is observed between impact score and LinkedIn measurement that assess social network influences. Such positive correlation between impact score and other measurements indirectly assessing impacts provides good evidence that RMDS's impact score is valid in measuring data scientists' impact.

The future update of our impact scores will consider including new features, fixing metrics issues, and seeing a slight overall increase in both validity and reliability of the scores.