6.S079 Project Report

Sophia Lockton
MIT
slockton@mit.edu

Ethan Ardolino
MIT
ardolino@mit.edu

Brady Klein
MIT
bmklein@mit.edu

ABSTRACT

Choosing the right college is a pivotal decision for high school students. U.S. News and World Report is considered by many to be the premier source for a comprehensive ranking of Universities. However, the methodology of U.S. news and the impact of rankings on admissions each year have garnered criticism. This project explores the intricacies of U.S. News' rankings by attempting to replicate the rankings based on a published methodology, analyzing correlations between rankings and admissions statistics, and testing the sensitivity of a comparable ranking algorithm. Utilizing comprehensive datasets from the Integrated Postsecondary Education Data System (IPEDS) and U.S. News, our approach examines many commonly asked questions about the U.S. News' rankings. How easy it is to game the ranking system, what impact do rankings have on admissions, and how sensitive is the ranking process to small changes in statistics?

1 INTRODUCTION

The U.S. News & World Report college rankings have long been a significant factor in shaping perceptions of higher education institutions in the United States. High school students, parents, and university administrators often look to these rankings for guidance and validation. The influence of these rankings is profound; they can affect application rates, fundraising efforts, and institutional reputation. However, the methodology behind these rankings has been a subject of debate and criticism, with questions about its transparency, susceptibility to manipulation, and overall impact on the higher education landscape. This project aims to delve into these aspects by attempting to replicate the U.S. News ranking system, analyzing the robustness of its methodology, and exploring its broader implications.

Our approach begins with an attempt to replicate the 2020 U.S. News National University Rankings using publicly available data. By reverse-engineering the ranking methodology, we aim to understand the factors that contribute most significantly to a university's rank. This involves meticulous data collection, preprocessing, and the application of statistical methods to model the ranking algorithm. Our sources of data include the Integrated Postsecondary Education Data System (IPEDS) and archival datasets that approximate the U.S. News methodology. This step is crucial as it lays the foundation for further analysis and critique.

To achieve this replication, we employed two primary methods: linear regression and weighted average calculations. Linear regression helps us understand the relationship between various ranking criteria and the overall scores, while weighted averages attempt to mimic the actual ranking process by assigning weights to different criteria as described by U.S. News. Both methods offer insights into the complexities and potential biases inherent in the ranking system. This process not only validates our data but also highlights any discrepancies between our results and the official rankings.

paragraph on conducting sensitivity testing

The historical context of these rankings is equally important. Over the years, U.S. News & World Report has been both lauded and criticized for its impact on higher education. Studies like those by Bastedo and Bowman [1] (2011) and Meredith [3] (2004) have examined how these rankings influence institutional behavior and student choices. Bastedo and Bowman highlight the significant organizational changes universities may undergo to improve their standings, while Meredith's work underscores the competitive nature of universities striving to climb the rankings. These studies provide a framework for understanding the historical and ongoing implications of the U.S. News rankings, reinforcing the need for a critical examination of their methodology and impact.

paragraph on observing historical implications of these rankings

In summary, this project not only seeks to replicate the U.S. News college rankings but also to critically evaluate their methodology and explore their broader implications. By doing so, we hope to contribute to the ongoing discourse on the validity and influence of college rankings in shaping higher education. Our findings aim to provide a more nuanced understanding of how these rankings are constructed and the extent to which they reflect the true quality of educational institutions.

2 DATA USED

U.S. News discusses the type of data that is used as well as the categories and respective weights of their ranking system. However, they do not publicly release the data associated with their yearly rankings. Understandable since the rankings are for-profit IP. Nonetheless, certainly frustrating for our project. Luckily, we were able to find archival data for this project from Prof. Andrew G. Reiter[4]. His personal website contains a relatively comprehensive approximation

of the U.S News College Rankings dataset. Specifically, we are using data relating to the 2020 U.S. National University Rankings. Additionally, we scraped an old revision history of Wikipedia to find the proper categories and weights for the 2020 U.S. News National Universities Ranking. We chose the year 2020 in that Prof. Reiter's dataset contains a majority of the relevant categories used in the 2020 rankings. The relevant columns used from Prof. Reiter's dataset that we are using is shown in Table 1. The categories and respective weights are for the 2020 rankings are shown in Table 2.

Table 1: Schema of 2020 U.S. News Dataset

Column Name	Description
rank	The rank of the university.
school	Name of the school (and state).
score	Overall score of the school calculated in U.S. News 2020 rankings
peer_score	Score based on the survey of institution's reputation among peers.
grad_rank	Rank based on graduation and first-year retention rates.
graduation_diff	A metric showing the difference between expected and actual graduation rates.
social_mobility_rank	Rank based on the school's contribution to social mobility.
faculty_rank	Rank based on faculty resources such as class size, faculty salary, etc.
test_scores	Average standardized test scores of admitted students.
impressive_freshman	A measure or score of the quality or achievements of incoming freshmen.
fin_resources_rank	Rank based on the financial resources per student related to academics, support, and public service.
alum_giving_rate	Rate of alumni donations.

A pitfall of this data is that the graduation/retention, social mobility, faculty resources, and financial resources columns all are displayed as a "rank". These categories make up 57% of the score within this ranking system, and thus not having the actual values is problematic. We discuss how we dealt with this in section 4.1.1.

Table 2: U.S News Best National Universities Ranking Categories and Weights

Category	Description	Weight
Peer assessment	A survey of the institution's reputation among presidents, provosts, and admissions deans of other institutions	20%
Retention	Six-year graduation rate and first-year student retention rate	22%
Social mobility	Six-year graduation rate of students receiving Pell Grants—both as a standalone measure and compared to graduation rates of all other students at the school—adjusted significantly to give more credit to schools enrolling larger proportions of students receiving Pell Grants	5%
Faculty resources	Class sizes, faculty salary, faculty degree level, student-faculty ratio, and proportion of full-time faculty	20%
Student excellence	Standardized test scores of admitted students and proportion of admitted students in upper percentiles of their high school class	10%
Financial resources	Per-student spending related to academics, student support and public service	10%
Graduation rate performance	Comparison between modeled expected and actual graduation rate	8%
Alumni giving rate		5%

3 MOTIVATION

Every year, high school students are heavily influenced on which colleges they apply to and attend by college rankings. U.S. News is the focus for this project due to its popularity among high school students. In 2022, a survey conducted over 800 graduating seniors showed that 58% "actively considered rankings in some way during their search", U.S. News being

used the most by roughly 22% of students[2]. Let that sink in. Nearly 1 in 4 high school students interacted with U.S. news while conducting their college search.

Further, the U.S. News rankings have received public scrutiny. First, the ranking methodology is simply stated and changed year over year. No rigorous explanation is given to elaborate on the rationale behind the choice of categories and weights that go into the scoring. Further, critics have discussed that the rankings can be "gamed". This is evident from Columbia University downgrading to rank 18 in 2023 after having been in the top 5 for 12 consecutive years due to false self-reporting. Additionally, it has been shown that some schools base their agendas off of these rankings. Rather than focus on concrete issues or shortcomings their specific school is facing, they worrying about the ranking on a list.

Point being, that given U.S. News' heavy influence on high school students and college administration, we wanted to assess the methodology and validity of these rankings.

4 SYSTEM DESIGN

The following three sub-sections will speak of the three distinct analysis projects that we conducted. First, we discuss creating a replication of the U.S. News National Universities Rankings for 2020. Next, we speak of the correlations that these rankings have with admission statistics. Lastly, we simulate adding gaussian noise to the criteria weights and select self-reported columns to assess the robustness of the ranking system.

4.1 Rank Replication

This section walks through the preprocessing steps made to clean and standardize the data shown in Table 1. We then discuss various attempts at replicating U.S News' 2020 National Universities rankings.

4.1.1 Preprocessing. The raw data columns were initially formatted as strings. Thus, after selecting the relevant columns shown in Table 1, we explicitly specified the data types for each column. This was trivial for *ints*, *floats*, and *strings* as these are supported directly by *pandas*. However, numerous columns had contextualized data such as *ratios* (i.e "7 to 4"), *percentages* (i.e "93%"), *ranges* (i.e "32-34"), and *ranks* that needed to be parsed and/or transformed.

We converted the *ratios* and *percentages* to their decimal representations. The only column that was represented as a *range* were ACT/SAT test scores. To standardize this column, we first converted the top and bottom of the *range* to their respective percentiles for the year 2020 and then took the simple average. We transformed all columns represented as a *rank* with the following equation:

$$r_i = \frac{max(R) - r_i + 1}{max(R)} \qquad \forall r_i \in R$$

This *rank* transformation is a bit naive; however, we felt that without any understanding of the underlying distributions of each column it was the most appropriate one to apply.

Lastly, we standardized each column using *z-scores* as U.S. News stated that they applied this to all of their collected data.

4.1.2 Linear Regression Replication. We began our analysis by using *sklearn*'s linear regression model. This was a quick and trivial way to assess the validity of our data. We regressed the reported scores against the ranking criteria. In an optimal setting, our model parameters would simply be the weights of the categories as the ranking system is described as a simple weighted average. There are nuances the normalizing and scaling the scores that are missed; however, we hoped that this was still the proper first step to take in our attempt to replicate the scores. We will discuss the performance of this model in section 5.1

4.1.3 Weighted Average Replication. First, we created a dictionary embodying the ranking categories and weights shown in 2. We then iteratively calculate a raw predicted score (column=p_score) for each row within our dataset. This is rather straightforward. The next stage of normalizing the p_score column is where a large amount of inference was made on our end. U.S. news states *"[t]he overall scores were scaled so that the top performer(s) in each ranking displays an overall score of 100. Others' overall scores are on a 0-99 scale reflecting the distance from their ranking's top-performing school(s)"*[5]. We perform the first step of normalizing so that $max(p\_score) = 100$. The notion of "distance from the top-performing school" is rather vague. Off of first principles, we selected to incorporate a linear transformation to all of p_score based off of the true reported scores (column=$score$). We could not scale the data linearly between 0-99 based off of the p_scores in that the dataset only contained the set of individually ranked schools (in 2020 the top 75% of ranking schools). Thus, we applied a min-max normalization based off of $min(score), max(score)$ to our predicted scores to accurately reflect the distance scaling method used in U.S. News. We will discuss the accuracy of our d_score in section 5.1

4.2 Noise Sensitivity

Given the difficulties we experienced when trying to replicate the U.S. News rankings, we wanted to explore how sensitive these rankings are to noise. To evaluate this, we first applied noise to the data columns and then to the weights used to determine the rankings (figure 2).

4.2.1 Adding Noise to the Data. In addition to our difficulty with replicating the rankings, the recent controversy

surrounding Columbia University's fraudulent data [6], we asked ourselves how much accuracy can we reasonably expect from universities on self-reported data? Or in the adversarial case, how sensitive might these rankings be to misreported data or schools that attempt to game the system?

When looking at the weights that U.S. News uses, we observed that 3 of the criteria were particularly susceptible to inaccurate reporting: faculty resources, financial resources, and alumni giving rate. To experiment how sensitive the ranking is, we decided to apply Gaussian noise to each of these self-reported columns and observe how the top-50 universities changed. We used Gaussian noise with standard deviations of 1%, 5%, and 10%.

After applying the noise to the columns, we evaluated the resulting dataframes by using Spearman's rank correlation coefficient. This metric is a non-parametric measure of the strength and direction of the association between two ranked variables. It evaluates how well the relationship between two variables can be described using a monotonic function. We conducted 1000 trials where we generated noise and applied it to the data, re-scored the schools, ranked them by these scores, and then calculated the correlation of this new ranking with the ground truth top-50. At the end of the 1000 trials, we took the mean of these correlations to produce an overall metric for how much the noisy data correlated with the ground truth on average when that level of noise was applied.

When applying only 1% noise on the data, we can see in figure 1 that the correlation to the ground truth top-50 did not change much. However, we see in figure 2 that when we apply 5% noise to the data we see significant drops in correlation across all 3 columns, proportional to their overall weight. When that noise increases to 10% as we see in figure 3, almost no correlation is present between the new rankings and the ground truth.

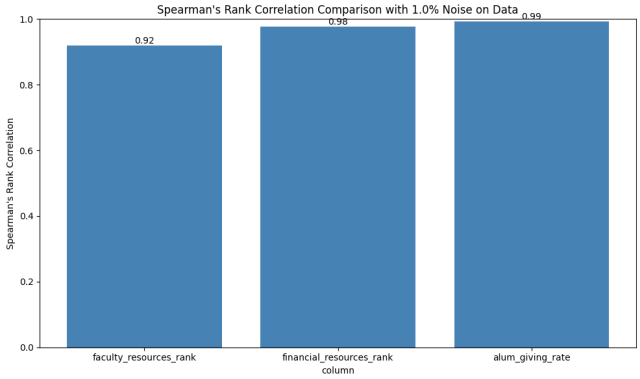

Figure 1: Spearman's Rank Correlation Coefficient with 1% Noise on Data

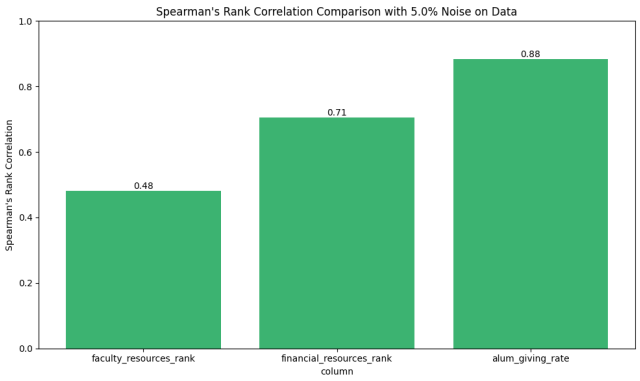

Figure 2: Spearman's Rank Correlation Coefficient with 5% Noise on Data

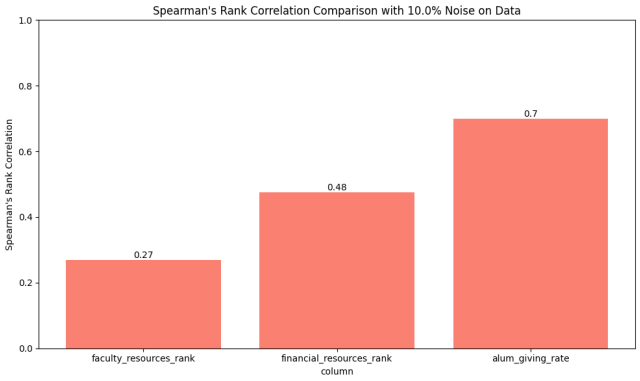

Figure 3: Spearman's Rank Correlation Coefficient with 10% Noise Levels on Data

4.2.2 Adding Noise to the Weights. After seeing how noisy data affected correlation with the ground truth top-50, we next wanted to analyze how applying noise to the weights themselves would affect the rankings.

For these purposes, we used the same setup of Gaussian weights, trials, and evaluating with Spearman's rank to determine the relationship. We can see these results in figure 4. As expected, the correlation dropped as the noise level increased. Interestingly enough, even at 10% noise, the correlation remained fairly strong.

4.3 Rankings and Admissions Statistics

5 EVALUATION

5.1 Replication

Below are four graphs showcasing the performances of the Regression and Weighted Average replication techniques. Figures 5 and 6 plot the reported rank of all schools against

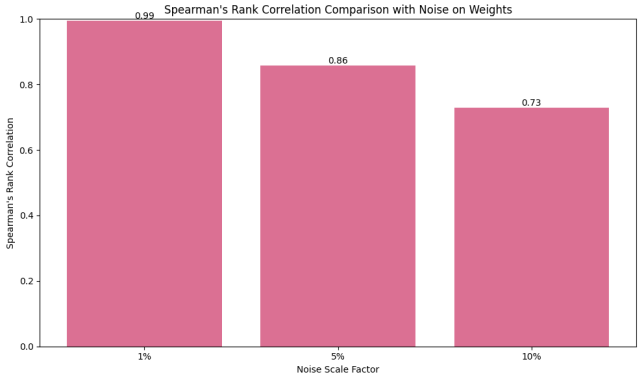

Figure 4: Spearman's Rank Correlation Coefficient with Various Noise on Weights

the reported and predicted score for both replication approaches. Visually, it is clear that the Weighted Average approach does not perform nearly as well as the Linear Regression technique. This discrepancy arises for two main reasons. First, our dataset embodies 57% of the criteria weighting across four categories as a *rank*. Thus, it is highly likely that our linear transformation of *ranks* as described in section 4.1.1 does not reflect the reported data accurately. Secondly, it appears that the majority of our predictions are greater than the reported scores. Apart from data quality, this could be caused from an improper distance scaling. Both of these concerns boil down to U.S. News lacing transparency in their data and methodology of their college rankings. It is a trade-off between protecting their IP (as they are a for-profit entity) against clarity in how the rankings are made.

Although the Linear Regression technique performs better, it certainly has its own flaws. The first issue are the model parameters shown in table 3. They are not representative of the criteria weights in the slightest. Trivially they are not identical due to the distance scaling; however, they are certainly not a linear transformation away from the criteria rankings either. Thus, although it performs relatively well, the model does not truly replicate the ranking system properly. Secondly, as can be seen in Figures 8 and 7, the Linear Regression replication performs significantly worse on the top 10 schools compared to the Weighted Average replication. This is likely due to the model overfitting to lower scores as they are more prevalent. This is important to us in that further analyses we wanted to conduct revolved around the higher ranked schools. In regards to the Weighted Average replication performing rather well on the top 10 schools, we do not have much insights. It is possible that Prof. Reiter's dataset has more accurate information pertaining to more prestigious schools. However, we do not have a compelling case to make this argument.

Table 3: Linear regression parameter values

Category	LR Coefficient
Peer Score	9.18
Graduation/retention	15.37
Graduation Rate performance	0.26
Social Mobility	4.013
Faculty Resources	8.94
Test Scores	0.025
Accomplished Freshman	10.6
Financial Resources	-0.015
Alumni Giving Rate	32.4

6 CONCLUSION TODO

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place.

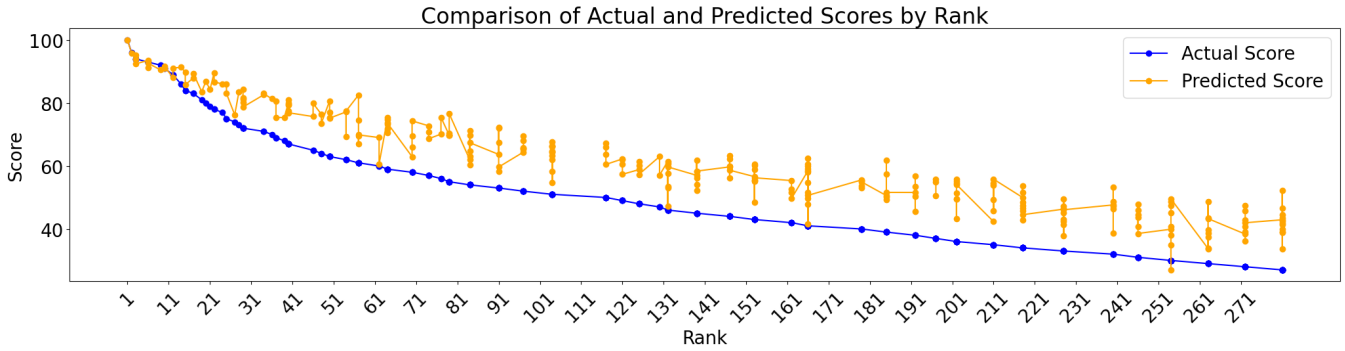

Figure 5: (MSE=170.82) Comparing the reported score and the predicted score using the Weighted Average replication technique for all ranked schools. Note that breaks in graph due to certain schools missing at least one of the weighted criteria.

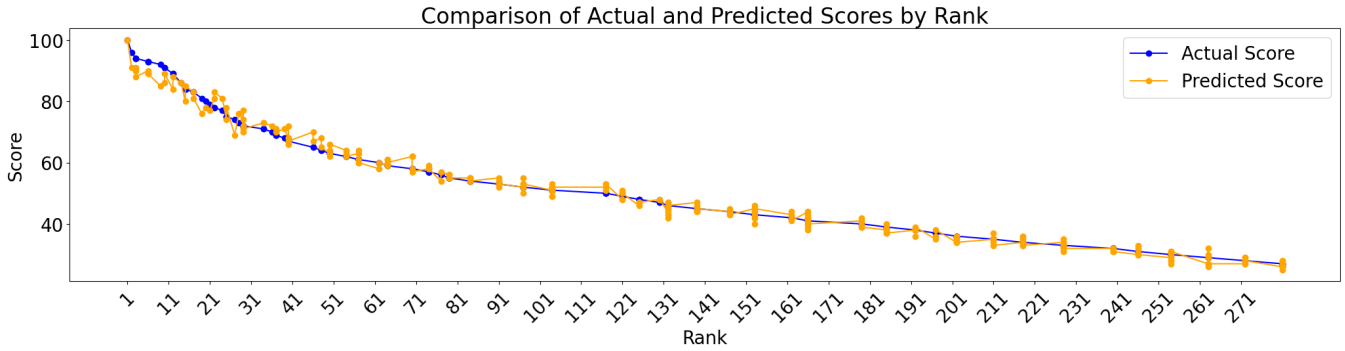

Figure 6: (MSE=3.34) Comparing the reported score and the predicted score using the Linear Regression replication technique for all ranked schools.

If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

REFERENCES

[1] Michael N. Bastedo and Nicholas A. Bowman. 2011. U.S. News & World Report college rankings: Modeling institutional effects on organizational reputation. *American Journal of Education* 117, 4 (2011), 563–587.

[2] Jessica Blake. 2023. U.S. News, not the behemoth one might think. *Inside Higher Ed* (13 September 2023). https://www.insidehighered.com/news/admissions/2023/09/13/us-news-not-behemoth-one-might-think

[3] Marc Meredith. 2004. Why do universities compete in the ranking game? An empirical analysis of the effects of the US News and World Report college rankings. *Research in Higher Education* 45, 5 (2004), 443–461.

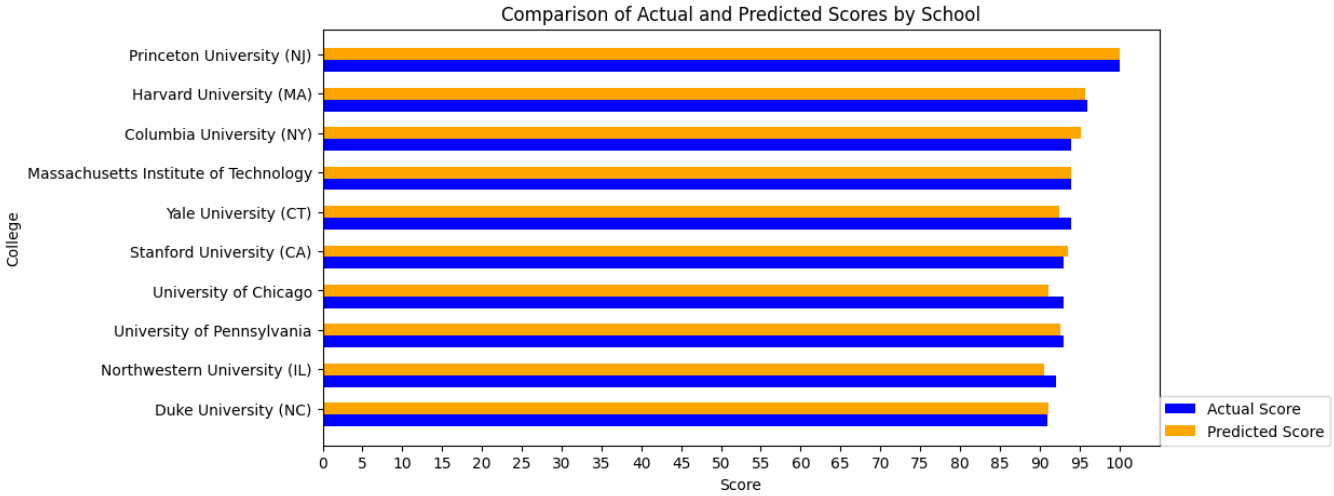

Figure 7: (MSE=0.99) Comparing the reported score and the predicted score using the Weighted Average replication technique on the top 10 schools.

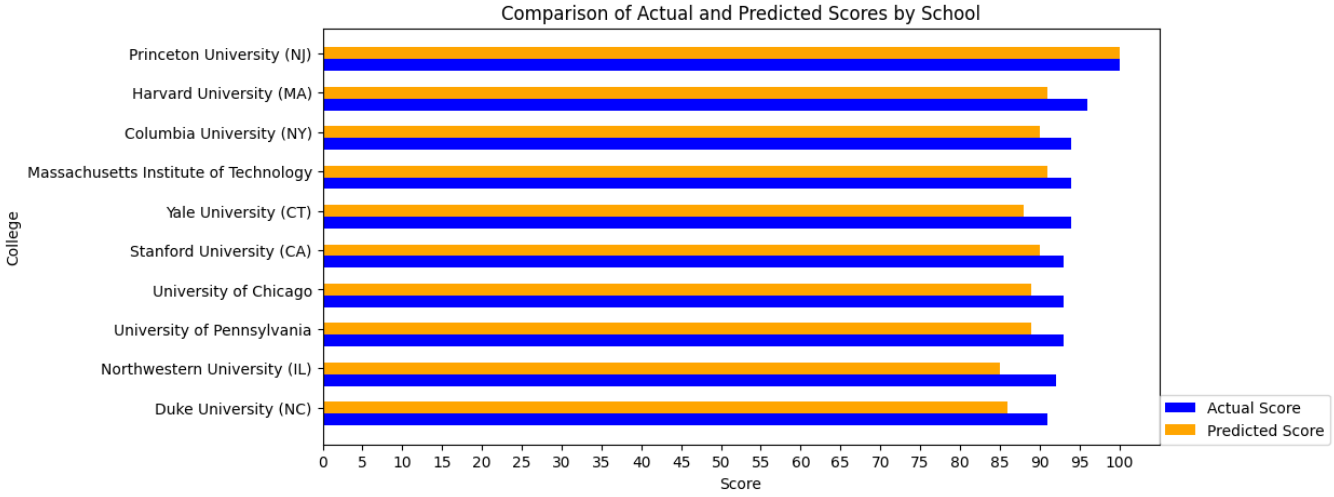

Figure 8: (MSE=20.10) Comparing the reported score and the predicted score using the Linear Regression replication technique on the top 10 schools.

[4] Andrew G. Reiter. 2023. U.S. News & World Report Historical Liberal Arts College and University Rankings. (2023). Available at: http://andyreiter.com/datasets/.

[5] Kevin Brooks Robert Morse. 2023. How U.S. News Calculated the Rankings. (2023). Available at: https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings.

[6] Alexandra Tremayne-Pengelly. 2022. Columbia Drops From No. 2 to No. 18 on University Rankings As School Officials Admit to Misleading Data. *Observer* (2022). https://observer.com/2022/09/columbia-drops-from-no-2-to-no-18-on-university-rankings-as-school-officials-admit-