**Competitive Rank Advantage: Elevating University Rankings through Strategic Analysis**

**Executive Summary:**

University rankings play a pivotal role in guiding prospective students to make informed decisions about their educational journey. These rankings, spanning various criteria such as subjects, continents, and countries, hold significant influence. This report concentrates on the comprehensive World University Rankings for 2023, aiming to unravel the multifaceted relationship between diverse factors and the overall university score.

Our focus lies in employing predictive models to recognize the pivotal factors influencing university rankings. By rigorously testing different models, this report seeks to pinpoint the critical elements that shape a university's overall score, thereby determining its ranking. This analysis strives to provide actionable insights, guiding Harvard University on strategic areas of improvement and offering a roadmap to ascend the global ranking scale.

This report thoroughly examines managerial implications stemming from the predictive analysis of world university rankings and also makes strategic recommendations. The report concludes in a comprehensive understanding of the influential factors, empowering Harvard to strategically enhance their positions on the global stage.

**Table of Contents**

**Introduction…………………………………………………………………………………5**

* + Background
  + Objectives

**Methodology…………………………………………………………………………………6**

**Data Collection and Wrangling…………………………………………………….6**

* + Dataset Description
  + Data Wrangling Techniques

**Exploratory Data Analysis (EDA)………………………………………………….7**

* + Analysis of Top 50 Universities' Locations
  + Top 10 Universities for Each Scoring Metric
  + Distribution of Scores within Overall Top 10 Universities

**Model Development and results……………………………………………………14**

* + Selection of Machine Learning Models
  + Model Comparison on Train and Test Datasets
  + Cross-Validation Results
  + Learning Curve Analysis
  + Extracting Feature Importance from Gradient Boosting Regression

**Conclusion…………………………………………………………………………………...19**

* + Summary of Findings

**Limitations and Further Research………………………………………………………...20**

**References…………………………………………………………………………………...21**

**Appendix…………………………………………………………………………………….23**

**Introduction**

As the educational landscape witnesses a surge in the number of universities and a continuous enhancement in the quantity and quality of learning, the competition among educational institutions has intensified significantly. In this dynamic environment, universities are compelled to make strategic decisions to carve their niche and provide cutting-edge, high-quality education. The pressing question for all universities resonates loudly: "Which areas should universities prioritize to improve in order to attract a greater number of students seeking to shape their future?"

Consequently, the pursuit of excellence becomes imperative for universities, driven by the ultimate business goal of maximizing profitability. The demand for admission to a university directly correlates with its potential for profit. In this context, a university's ranking emerges as a pivotal factor in attracting students, securing funds for scientific endeavours, garnering recognition through awards, contributions to significant discoveries and for dynamic as well as continuous improvement (Vidal, J. and Ferreira, C., 2020).

At the forefront of the educational institution landscape stands Harvard University, a beacon of leadership and innovation. Positioned at the second rank in the ‘World University Rankings 2023’, Harvard aspires to rise to the coveted top spot. This report, rooted in Harvard's ambition, aims to unravel the key aspects Harvard must enhance to claim the leadership position.

Within the context of the management strategy vertical at Harvard University, a precise examination of each influential factor provides a comprehensive understanding of the university's strengths and areas for improvement. This report offers profound insights and actionable strategies to propel Harvard University to its desired leadership position. Leveraging and comparing predictive supervised machine learning models—namely Decision Tree, Random Forest, and Gradient Boosting Regression— the report dissects influential factors and assesses their individual importance in determining the overall score (a critical metric that intricately shapes the ranking of any university) for any university.

The ensuing discussion provides the rationale behind choosing these predictive models, evaluates their efficacy, and delivers valuable insights into selecting the optimal model. Additionally, limitations and avenues for improvement are explored, ensuring a holistic view of the predictive analysis landscape.

**Methodology**

**Data Collection and Wrangling**

To accomplish the objectives outlined in this project, we leverage the 'World University Rankings 2023' dataset, a valuable open-source resource sourced from Kaggle. This comprehensive dataset encompasses information on 2233 universities spanning 118 countries and regions. The dataset is meticulously curated, incorporating 13 well-calibrated performance indicators (Refer to **Appendix**) designed to assess an institution's performance across four critical domains: teaching, research, knowledge transfer, and international outlook. The dataset provides a robust foundation for our analytical endeavours, offering insights into the diverse landscape of universities worldwide.

Following the import of the 'World University Rankings 2023' dataset into our Kernel, a comprehensive data wrangling and pre-processing phase is undertaken to prepare the dataset for subsequent predictive modelling. The initial examination reveals the presence of NaN values in multiple rows and columns, necessitating meticulous handling.

Firstly, rows lacking the name of the university are identified and subsequently removed from consideration. Moving on to the 'Location' column, missing values are addressed using the Nominatim function from the geopy library in Python. For the 'Female:Male Ratio' column, it is observed that the current format is not favourable to our modelling approach. To address this, the column is split, and percentages for females and males are extracted into distinct columns. Subsequently, a new feature is also engineered —‘abs Male’— which depicts the gender equality in a university. It is formulated by subtracting 50 from the ‘Male’ column (which depicts the % of males in the university).

The 'International Student' column, which represents the percentage of international students, is hindered by the presence of '%' signs, causing it to be treated as an object rather than a numeric variable. The '%' signs are removed, and the percentages are rescaled between 0-1.

Turning attention to the 'OverAll Score' column, certain rows present a range rather than an exact value, particularly in lower-ranked universities with minor distinctions. In these cases, the maximum value within the range is considered, given the negligible impact of using the averages.

For the 'No of Student' column, the presence of commas in the numeric scale results in the column being treated as an object. Removal of the commas and parsing the column as a numeric variable resolved this issue.

Finally, remaining rows and columns containing NaN values, including 'No of student', 'International Student', 'Female', 'Male', 'OverAll Score', 'No of student per staff', 'Teaching Score', 'Research Score', 'Citations Score', 'Industry Income Score', and 'International Outlook Score,' are handled by replacing them with the mean of all values within their respective columns. This thorough data wrangling process ensures the dataset is refined and primed for subsequent predictive modelling.

**Exploratory Data Analysis and Summary**

During the exploration of our selected dataset, we uncovered several intriguing insights that laid the foundation for our predictive analysis. Examining the geographic distribution of the top 50 universities, we observed that nearly 50% of these institutions are located in the United States, with Harvard University being a prominent example. Although location doesn't emerge as a significant factor affecting rankings, the trust associated with the locality remains a crucial subjective consideration, an aspect where Harvard excels.

**Figure 1: Top 50 Universities Locations**

A map of the world with different colored points

Description automatically generated

Subsequently, we delved into the top 10 universities for various scoring metrics, including Teaching, Research, Citations, Industry Income, and International Outlook. We also explored the distribution of these scores within the overall top 10 ranked universities.

***Teaching Score***

Diving into the Teaching Score metric, Harvard's prominence was evident as it bagged the top rank. This achievement underscores an area where Harvard excels—teaching quality. The university should aim not only to sustain but also to augment this strength in the future to fortify its overall position. Intriguingly, the absence of certain universities from the overall top 10 in this metric implies that only seven institutions (Harvard’s direct competitors) pose a challenge to Harvard's comprehensive ranking with this score.

**Figure 2 : Top 10 Universities for Teaching Score and the Distribution of Teaching Score across the Overall Top 10 Universities of the World**

A grey rectangular background with white text

Description automatically generatedA graph with blue and white stripes

Description automatically generated

A table with a list of college names

Description automatically generated with medium confidenceA graph with blue and white stripes

Description automatically generated

***Research Score***

The Research Score metric disclosed that Harvard secured the 3rd rank globally. To bolster its standing in this domain, Harvard must contemplate the allocation of additional resources towards bolstering research quality and quantity. Disparities in scores between Harvard and the 4th-ranked university, and the nearness of scores between Harvard and the top 2 ranked universities emphasize the imperative for relentless efforts to achieve research goals without worrying about competitive pressures.

**Figure 3 : Top 10 Universities for Research Score and the Distribution of Research Score across the Overall Top 10 Universities of the World**

A graph with blue and white stripes

Description automatically generatedA black and white screen with numbers

Description automatically generated with medium confidence

A bar graph with blue and white stripes

Description automatically generated

A table with text on it

Description automatically generated

***Citations Score***

A detailed analysis of the Citations Score revealed an intriguing pattern: none of the overall top 10 global universities made it to the top 10 list for this metric. Harvard, however, clinched the 3rd position in the distribution of this score among the overall top 10. Given the intensely competitive landscape (Refer to Figure 4) amongst the overall top 10 ranked universities, Harvard must continuously enhance the usability and relevance of its research work to ascend higher in this metric, which signifies the global impact of scholarly contributions.

**Figure 4 : Top 10 Universities for Citations Score and the Distribution of Citations Score across the Overall Top 10 Universities of the World**

A graph with blue and white lines

Description automatically generatedA screenshot of a cellphone

Description automatically generated

A table with text on it

Description automatically generatedA bar graph with blue and white stripes

Description automatically generated

***Industry Income Score***

Our exploration of the Industry Income Score showcased a trend similar to the Citations Score: none of the top 10 global universities featured in the top 10 list for this metric as well. Harvard, unfortunately, secured the last position in the distribution of this score among the overall top 10, signalling a notable gap. To bridge this divide, Harvard needs to commit substantial resources, concentrating on bolstering the industry presence and relevance of its research endeavours. This metric is intricately linked to both citations and research scores, highlighting a compelling need for a concentrated emphasis on enhancing research-based endeavours.

**Figure 5 : Top 10 Universities for Industry Income Score and the Distribution of Industry Income Score across the Overall Top 10 Universities of the World**

A grey rectangular background with white text

Description automatically generatedA graph with blue and white stripes

Description automatically generated

A grey rectangular table with white text

Description automatically generatedA bar graph with blue and white stripes

Description automatically generated

***International Outlook Score***

Within the International Outlook Score, none of the overall top 10 universities globally found a place in the top 10 list for this metric (same as the last two scoring metrics). Harvard secured the 6th position in the distribution of this score among the overall top 10. Enhancing internationalization efforts, promoting diversity, and fostering global collaborations are pivotal strategies for Harvard to excel in this metric, which gauges the university's global relevance and engagement.

**Figure 6 : Top 10 Universities for International Outlook Score and the Distribution of International Outlook Score across the Overall Top 10 Universities of the World**A black and white text on a grey background

Description automatically generatedA graph with blue and white stripes

Description automatically generated

A black and white text on a grey background

Description automatically generatedA graph with blue and white stripes

Description automatically generated

***Gender Equality Score***

The Gender Equality Score underscored another triumph for Harvard as it emerged as the leader. This competitive advantage in gender equality is significant not only from a moral standpoint but also because gender equality within educational institutions is recognized as a catalyst for economic growth (Bertay, A.C., Dordevic, L. and Sever, C. (2020)). Harvard's continuous commitment to upholding gender equality among students and staff positions aid them to be a leader in fostering an inclusive and equitable learning environment.

**Figure 7 : Top 10 Universities for Gender Equality Score and the Distribution of Gender Equality Score across the Overall Top 10 Universities of the World**

A screen shot of a black and white screen

Description automatically generatedA graph with numbers and symbols

Description automatically generated

A screen shot of a number of college students

Description automatically generatedA graph with blue and white bars

Description automatically generated

**Model Development and Results**

The primary aim of this report is to assess and communicate the significance of various factors influencing a university's global ranking to Harvard University. This involves employing supervised machine learning models to quantify feature importance, enabling the identification and documentation of potential shortcomings in Harvard University's attributes, preventing it from securing the desired top-ranking position and currently holding the 2nd spot globally. Throughout this analysis, three distinct supervised machine learning models have been explored and evaluated: Decision Tree, Random Forest, and Gradient Boosting Regression.

The preference for employing non-linear regression models over linear regression models stems from the greater difficulty in interpreting feature importance in the latter. Non-linear models offer the advantage of constructing and comprehending intricate relationships between features and the target variable. Given the inherent noise in the data and the intricate nature of connections, such as the university's name, location, or gender equality, impacting the overall ranking score, the decision to utilize non-linear models appears appropriate (Saarela, M. and Jauhiainen, S. (2021)).

A Decision Tree is a tree-like model where each internal node represents a decision based on a feature, each branch represents an outcome of the decision, and each leaf node represents the final predicted label or value. A Random Forest is an ensemble method that builds multiple Decision Trees that merges their predictions, and Gradient Boosting Regression is another ensemble method, but it builds trees sequentially.

Opting for such non-linear models provides benefits in terms of enhanced interpretability and the model's capacity to learn from prior errors, thereby improving its output.

To conduct a comprehensive comparison of these models, the initial step involves pre-processing the data to prepare it for the prediction process. This is done by employing One-Hot Encoding for categorical variables and subsequently partitioning the dataset into training and testing sets.

Upon concurrently executing the three models and assessing their predictive performance on both the training and test datasets, noteworthy insights appear. On checking with the training set, the decision tree model exhibits a perfect R2 value of 1, while the random forest and gradient boosting models show R2 values of 0.9978 and 0.9934, respectively. This contrast clearly demonstrates that the decision tree model may not be well-suited for this analysis, as the R2 value of 1 indicates overfitting, underscoring the model's struggle to generalize to unseen data (Refer to Figure 8).

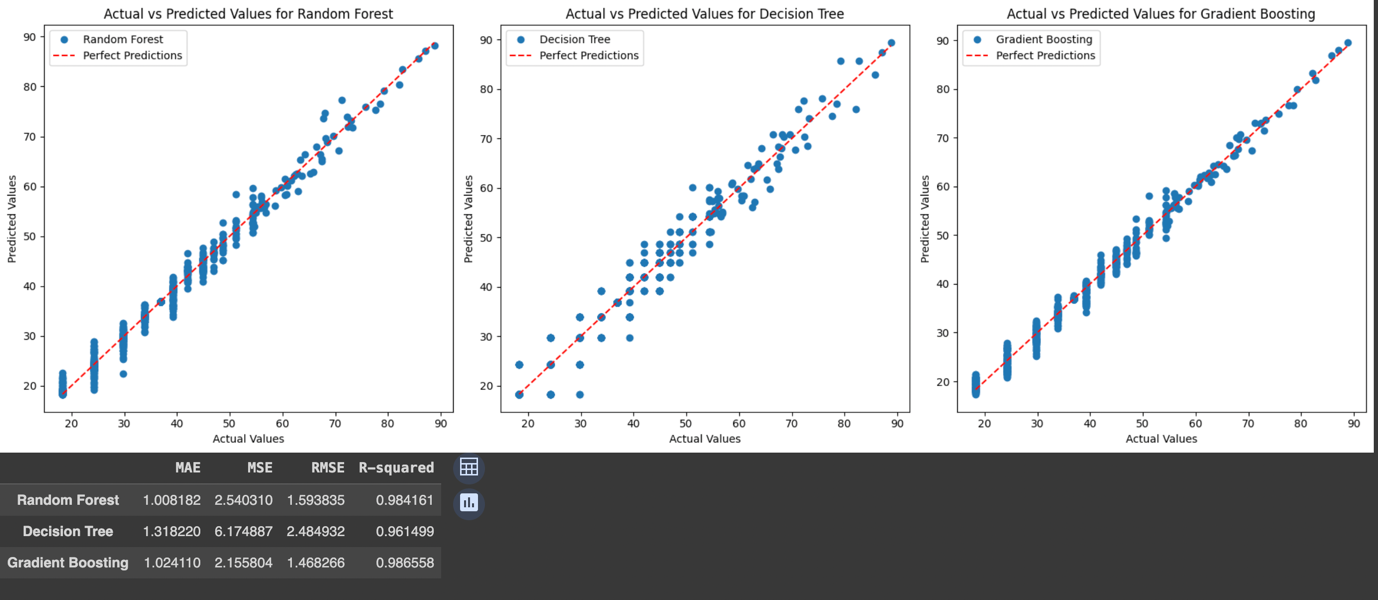
**Figure 8. Results for the 3 models on the *train* dataset**

**A screenshot of a computer

Description automatically generated**

Subsequently, upon examining the test set outcomes, we observe R2 values of 0.9841 and 0.9865 for Random Forest and Gradient Boosting, respectively. In contrast, the Decision Tree model yields an R2 value of 0.9614, reinforcing our earlier assertion of it not being a good fit (Refer to Figure 9).

**Figure 9. Results for the 3 models on the *test* dataset**



Before hastily concluding that Gradient Boosting is the optimal model, we perform a crucial cross-validation test on all three models to validate our initial findings.

Cross-validation, a pivotal technique in machine learning, involves dividing a dataset for training and testing, enhancing a model's generalization to new data points. It helps identify and address issues like overfitting, ensuring the model's predictive capability extends beyond sample data (Berrar, D. (2019)).

In the re-evaluation using cross-validation, Gradient Boosting Regression exhibits the lowest mean MSE and RMSE (2.125 and 1.456, respectively) and the highest Mean R2 (0.9877, i.e., 98.77% fit) across all three models tested proving its predictive ability is the strongest by far (Refer to Figure 10).

A screenshot of a computer

Description automatically generated**Figure 10. Results for cross-validation of all 3 models**

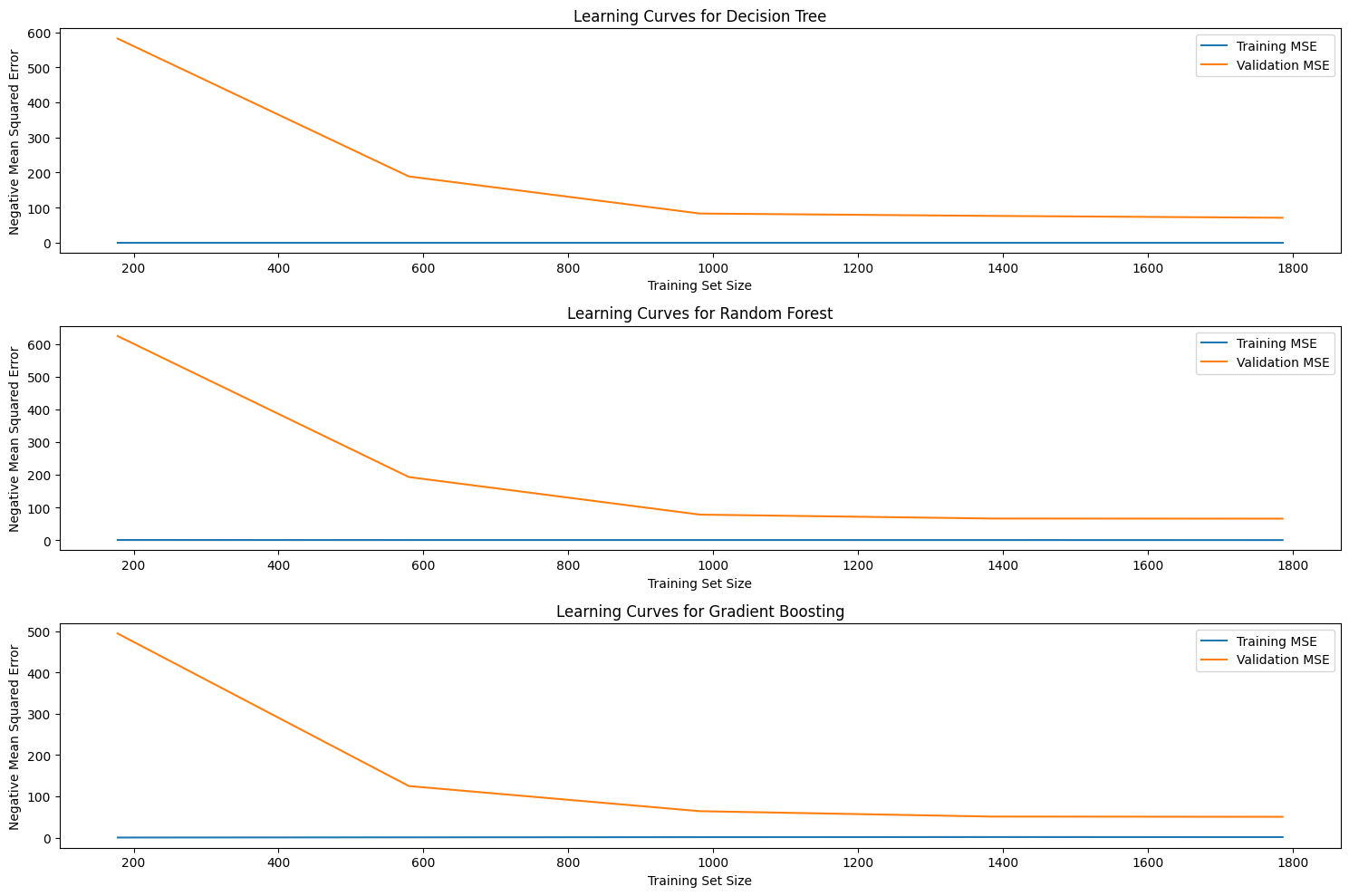
A graph of a graph with numbers and lines

Description automatically generated with medium confidence

Subsequently, by utilizing a learning curve to elucidate the performance and behaviour of each model as the training set size expands, it is shown effectively which particular model can predict novel and unseen data. The outcomes of the learning curve affirm that the learning ability of the Gradient Boosting Regression model surpasses that of both the other models. Gradient Boosting Regression exhibits a faster learning rate as shown by the slope of the Validation curve. As the validation curve converges with the training curve yet remaining parallel, it strongly suggests that the model has reached a point where additional information won't significantly enhance its learning, indicating readiness to predict unseen datasets (refer to Figure 11).

Considering this comprehensive and clear evidence, it is confidently concluded that Gradient Boosting Regression aligns perfectly with our objectives.

**Figure 11. Learning curves for all 3 models**

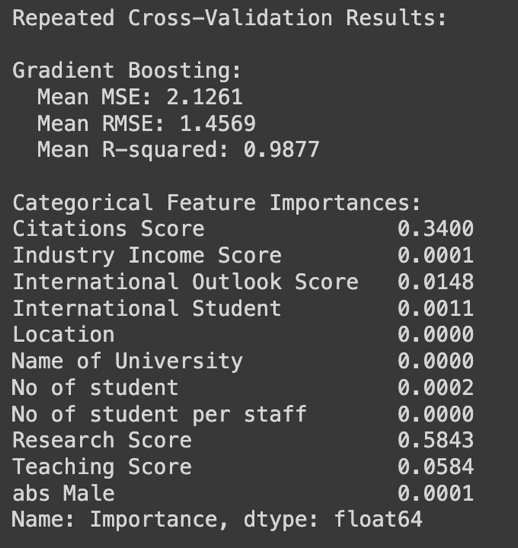


After selecting Gradient Boosting Regression as our predictive model, we proceeded to analyse feature importance to discern the impact of each factor on the overall university score. Given the extensive range of countries and universities in the dataset, conducting an in-depth examination of the individual effect of each encoded categorical variable might be impractical. Consequently, we opted for a practical approach by averaging out the importance of all individually encoded variables in a particular group, thereby revealing the significance of University Name and Location.

The outcomes present a clear yet intriguing insight into the determinants of a university's overall score. Notably, the **Citations Score** and **Research Score** emerge as the most influential factors in shaping the overall score of any university. In contrast, the **Teaching Score** demonstrates a comparatively lower importance as a feature in determining the overall score (refer to Figure 12). An additional insightful observation pertains to the comparatively modest impact of **Industry Income** and **International Outlook** Scores on the Overall Score of universities. Moreover, it becomes evident that both **categorical variables** and **engineered features** exert only a marginal influence on the Overall Score, underscoring their relatively limited contribution to the comprehensive assessment of features affecting university rankings.

Thus, a nuanced understanding highlights the dominant role played by **only** specific key factors, such as Citations Score and Research Score, in shaping the overall assessment of universities.

**Figure 12. Results for Repeated and Grouped Cross-Validation Feature Importance**



1. **Conclusion**
2. Bottom of Form

**Conclusion**

In summary, upon conducting a comprehensive analysis, a distinct and brief conclusion emerges. The pivotal factors influencing the Overall Score, namely Citations and Research Scores, reveal areas where Harvard University experiences a relative lag. In contrast, the Teaching Score, while not highly significant, stands as a strength for Harvard, showcasing leadership in this particular dimension.

The data underscores a compelling need for strategic improvements in the more influential scores—Citations and Research. Addressing these areas requires a multifaceted approach. Enhancing the Research Score involves a strategic allocation of resources, with a higher budget dedicated to research purposes. Additionally, encouraging faculty and students to actively engage in research activities and publish their work more frequently can elevate the university's standing in the academic community.

Considering the symbiotic relationship between Citations and Research Scores, it becomes evident that the qualitative enhancement of research endeavours directly correlates with increased citations and broader influence on a global scale. Thus, investing in and promoting impactful research can significantly augment Harvard University's relevance in contemporary academia, finally improving the Citations Score.

While maintaining its leadership position in Teaching Score remains imperative, the gradual improvement of the International Outlook score is also identified as a potential area of enhancement. This, however, can be pursued progressively, given its comparatively lower impact on the Overall Score.

In essence, the conclusion emphasizes a strategic focus on uplifting the research profile of Harvard University as a pivotal pathway to bolster its overall ranking and becoming the leader in the academic world.

**Limitations and Further Research**

Anticipating university rankings poses considerable challenges. These rankings resemble moving targets, influenced by a variety of ever-changing factors. Economic shifts, major global events, and evolving policies contribute to the dynamic nature of this ranking rollercoaster. Our predictive model faces a difficult task in capturing these fluctuations using historical data. Adding complexity, there's the subjectivity in rankings. Different organizations build their unique criteria and approaches, making comparisons relate to apples and oranges. Our model struggles with quantifying this subjective realm. It also has a blind spot for external influences that can disrupt rankings – factors like political stability, social dynamics, and the perceptions about faculty and institutions. These factors act as unpredictable variables challenging the model. While our model diligently seeks trends in this sample, exploring multiple ranking scales could enhance its ability to capture a broader array of features and offer a more comprehensive understanding of ranking scale feature importance.

**Total Words** - 3188

**References**

1. Bertay, A.C., Dordevic, L., and Sever, C. (2020). 'Gender Inequality and Economic Growth: Evidence from Industry-Level Data.' \*Social Science Research Network [Preprint]\*. [https://doi.org/10.2139/ssrn.3658594](https://doi.org/10.2139/ssrn.3658594).

2. Berrar, D. (2019). 'Cross-Validation.' In \*Elsevier eBooks\*, pp. 542–545. [https://doi.org/10.1016/b978-0-12-809633-8.20349-x](https://doi.org/10.1016/b978-0-12-809633-8.20349-x).

3. Butler, N., Delaney, H., & Śliwa, M. (2017). 'The labour of academia.' \*ephemera\*, 17(3).

4. Fernández Tuesta, E., Garcia-Zorita, C., Rosario, R., & Sanz-Casado, E. (2019). 'Does a Country/Region’s Economic Status Affect Its Universities’ Presence in International Rankings?' \*Journal of Data and Information Science\*, 4(2), 56–78. [https://doi.org/10.2478/jdis-2019-0009](https://doi.org/10.2478/jdis-2019-0009).

5. García, J.V., and Ferreira, C. (2020). 'Universities Under Pressure: the Impact of International University Rankings.' \*Journal of New Approaches in Educational Research\*, 9(2), p. 181. [https://doi.org/10.7821/naer.2020.7.475](https://doi.org/10.7821/naer.2020.7.475).

6. Hazelkorn, E., and Mihut, G. (eds) (2021). \*Research handbook on university rankings: theory, methodology, influence and impact\*. Cheltenham, Gloucestershire: Edward Elgar Publishing Limited (Elgar Handbooks in Education).

7. Loyola-Gonzalez, O. et al. (2020). 'A Contrast Pattern-Based Scientometric Study of the Qs World University Ranking.' \*IEEE Access, 8\*. [https://doi.org/10.1109/ACCESS.2020.3037665](https://doi.org/10.1109/ACCESS.2020.3037665).

8. Rugangazi, B., Okeyo, G., and 2023 IEEE AFRICON Nairobi, Kenya 2023 Sept. 20 - 2023 Sept. 22 (2023). “2023 Ieee Africon,” in \*Detecting Phishing Attacks Using Feature Importance-Based Machine Learning Approach\*. IEEE, pp. 1–6. [https://doi.org/10.1109/AFRICON55910.2023.10293475](https://doi.org/10.1109/AFRICON55910.2023.10293475).

9. Saarela, M., and Jauhiainen, S. (2021). 'Comparison of Feature Importance Measures As Explanations for Classification Models.' \*SN Applied Sciences, 3(2)\*. [https://doi.org/10.1007/s42452-021-04148-9](https://doi.org/10.1007/s42452-021-04148-9).

10. Vidal, J., and Ferreira, C. (2020). “Universities Under Pressure: The Impact of International University Rankings.” \*Journal of New Approaches in Educational Research, 9(2)\*, pp. 181–193.

**Appendix**

|  |  |  |
| --- | --- | --- |
| No. | Field | Description |
| 1. | University Rank | Ranking of the university in 2023 |
| 2. | Name of University | The name of the university |
| 3. | Location | The country it is based in |
| 4. | No of student | The number of students studying in it in 2023 |
| 5. | No of student per staff | The number of students divided by the total quantity of staff |
| 6. | International Student | The percentage of international students in the university |
| 7. | Female:Male Ratio | The ratio of females to males in the university |
| 8. | OverAll Score | The overall score out of 100 (Based on the bottom 5 scores) |
| 9. | Teaching Score | The teaching score out of 100 (How efficient are university classes and staff?) |
| 10. | Research Score | The research score out of 100 (Is research performed here considered relevant?) |
| 11. | Citations Score | The citations score out of 100 (How often are authors from here cited in books or articles?) |
| 12. | International Outlook Score | The international outlook score out of 100 (Is the university relevant to the international community?) |
| 13. | Industry Income Score | The industry income score out of 100 (Do businesses give credit to the university?) |