An Analysis of Time Higher Education World University Ranking: A Visual Analytics Approach

Ho Yin Tam
Valli Ramaswamy
School of Science and Technology (SST)
Department of Computer Science
City, University of London

Abstract - With the emphasis on global education, the university ranking has become a vital reference for teaching quality, research contributions, citations impact, industry income, and international outlook to individuals. This paper will exploit the university ranking reports published by The Time Higher Education (THE) World University Ranking from 2011 to 2023, and adopt a visual analytics approach analysing the geographical distribution, subject distribution, and overall scoring distribution of the top 200 universities. Also, this paper will perform time series analysis on the ranking change among the top 10 universities from 2011 to 2023. In addition, this paper will investigate the correlation between the overall scores and the underlying factors and build the predictive models for predicting the overall scores.

1. Problem Statement

In modern society, higher education plays a significant role in developing critical thinking and helping individuals meet job qualifications. Individuals studying in a university with a higher ranking may have more academic resources, access to influential networks, and gain more job opportunities. The Time Higher Education (THE) World University Ranking, a widely recognized benchmark, provides a comprehensive overview of teaching quality, research contributions, citations impact, industry outcome, and international outlook, to prospective students. This research primarily uses the university ranking reports published by THE and focuses on answering the following questions:

1. How do geographical distribution, subject distribution, and overall scoring distribution of the top 200 universities compare in the years 2011 and 2023?
2. How did the ranking change among the top 10 universities from 2011 to 2023?
3. What are the factors determining the overall scores and how are they correlated?
4. Using the multiple linear regression model and random forest model, can the models accurately predict overall scores?

The data is suitable for answering the research questions as this dataset contains the teaching scores, international outlook scores, industry income scores, research scores, and citation scores of different universities, which are the indicators impacting the overall scores and in turn, the ranking of the universities. Also, since the report is published annually, combining several reports across a period of time allows for a time series analysis of university ranking. In addition, the data type is numerical, which is suitable for investigating the correlation between various attributes and constructing predictive models.

2. State of the Art

The three academic papers analysed provide critical thinking and inspiration for this research paper. The first paper ‘What contributes more to the ranking of higher education institutions? A comparison of three world university rankings’ published by Hou and Jacob compared three different global university rankings, including the QS World University Ranking, the Academic Ranking of World Universities (ARWU), and the Times Higher Education (THE), and examined the effects of the indicators on the overall ranking of the universities [1]. The paper used the dataset and the reports published by these three institutions from 2013 to 2014 to analyse the correlation between the overall rankings and their indicators of the top 100 universities using the bivariate regression model and the multiple linear

regression model. However, the bias occurs since the variation of the top 100 universities is smaller and the top 100 universities may change each year. Hence, in this paper, a total of 13 reports are combined and the top 200 universities in each report are chosen as the final dataset to be analysed to eliminate the bias.

The second paper titled 'A data analytics approach for university competitiveness: the QS world university rankings' [2] published by Ana Carmen Estrada-Real & Francisco J. Cantu-Ortiz presents how well a predictive model performs for measuring university performance in the QS world university rankings spanning across ten years. The time range and dimensions of the dataset are similar to the Times Higher Education (THE) university rankings dataset used in this research paper which adds validity to the predictive models to be developed for this research. The paper compares the performance of a Fixed Effects regression model and a Random Forest which is ideal as this paper would compare the performance of a Random Forest too however, since their Fixed Effects model did not prove effective, a multiple linear regression model will be chosen to be implemented.

The final paper, 'University Ranking Prediction System by Analysing Influential Global Performance Indicators' [3] by Anika Tabassum et al., focuses particularly on feature selection and exploratory data analysis. It places a strong emphasis on understanding the impact of performance indicators and developing predictive scoring models, closely aligning with the intended methodology for this research. The machine learning framework employed in the paper is consistent with the aims of this study, where the researchers aim to identify the actual effects of performance indicators and pinpoint the most influential factors. Their approach involves developing models for predicting scoring and subsequently evaluating their accuracy. One notable aspect shared between this paper and the following research is the initial steps taken in the analysis process. Both studies advocate for the creation of a correlation matrix and the assessment of feature distributions before building prediction models.

3. PROPERTIES OF THE DATA

The Time Higher Education (THE) World University Ranking Report is published annually, aiming to offer trustworthy university-related data to different stakeholders, such as students, their parents, the government, and the industries since 2004 [4]. In this paper, a total of 13 datasets collected in Kaggle are used for the formation of the final dataset and analysis of the university ranking from the year 2011 to 2023 [5]. Each dataset captures the world university ranking with respect to teaching quality, research contributions, citations impact, industry outcome, and international outlook.

The dataset does require a substantial amount of manual cleaning as each file consists of a different number of rows and columns which leads to a difficulty when concatenating. For example, a total of 200 rows and 20 columns are found in the '2011_ranking' file, while a total of 800 rows and 24 columns are found in the '2016_ranking' file. Hence, the first 200 rows are extracted for each file, representing the top 200 universities in the world. Also, the column of 'female and male ratio', 'percentage of international students', 'number of full-time equivalent students', and 'number of students per staff' are dropped as only some files contain these four columns. In addition, a new column 'Year' is added as no columns indicate the year in each file. Furthermore, a new column 'Region' is added to group the countries into regions, such as Japan being categorised as Asia, Australia as Oceania, and Egypt as Africa, etc. Then, the data frame is saved as a CSV file for further cleaning.

Once saved, certain columns require consistency, for example, the '=' sign or character is removed so the 'rank' column is transformed from 'object' data type to 'numeric' data type. To handle the missing values, the function 'isnull' in Python is used and only 16 missing values are found in the column 'subjects_offered'. Then, the 'info' function is used for checking the data type of each column. The 'scores of international outlooks' and 'scores of industry income' are 'object' data type, and hence they are changed to 'numeric' data type. After changing, a total of 135 missing values were found and removed.

Box-and-whisker plots serve as a diagnostic tool for identifying outliers which are the observations that deviate significantly from the majority of the data [6]. Figure 1 shows the box plots of the overall scores, teaching scores, international outlook scores, industry income scores, research scores, and citation scores. Notably, the citation scores have the highest occurrence of outliers. However, since the scores range from 0 to 100, a low or high score does not indicate an extreme value but rather provides valuable insights into the overall score. Therefore, it is reasonable to retain these observations instead of discarding them.

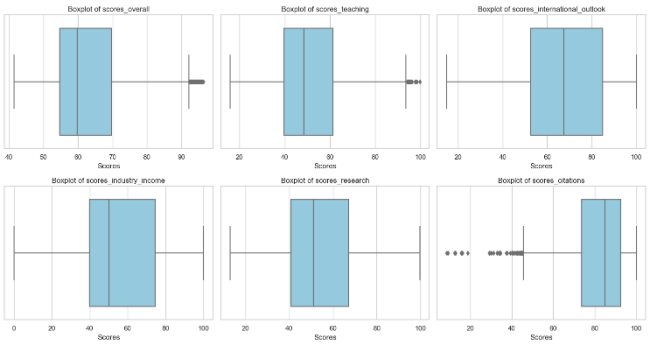

Figure 1: Box plots of chosen features

Figure 2 depicts the histogram showcasing the distribution of overall scores, teaching scores, international outlook scores, industry income scores, research scores, and citation scores. The overall scores and teaching scores tend to have a right-skewed distribution while the citation scores exhibit a left-skewed distribution.

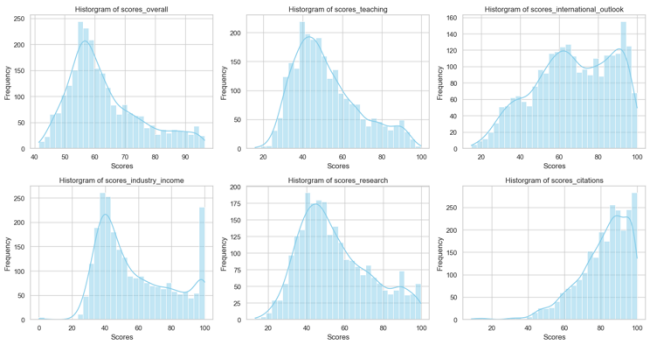

Figure 2: Histogram distribution of features

4. ANALYSIS

4.1 APPROACH

This section elaborates on the tasks undertaken to address our research questions. This diagram depicts the visual analytics process [7] Daniel Keim developed to approach a data analysis problem.

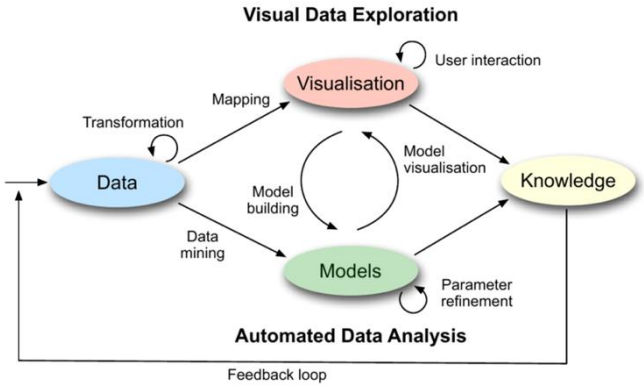

Figure 3: Visualisation Analysis Workflow (Keim et al, 2010)

The section below explores how each task, addressing its respective research question, is to be approached using Keim's workflow.

Task 1

Computation

For this task, three visualisations will be created but the dataset will first be pre-processed by filtering for specific years (2011 and 2023), and locations will be grouped by their specific geographical regions as human analysts will be able to easily digest regionally mapped data [8]. Statistical distributions will be computed in a box plot and the pie chart computation will follow similar pre-processing principles. For the bar chart, computer operations are required to split the string within the column 'subjects_offered' that will count each subject's occurrences to identify the top 10 most and least popular subjects in 2011 and 2023.

Visual

Both box plots and pie charts facilitate a human-centric approach as analysts can conduct cross-regional examinations and trend identification [9]. Visually the box plot communicates scoring distributions, its central tendencies, skews, and potential outliers whereas

pie charts offer a different slice of the data prompting users to assess significant changes over the 12-year period. The bar charts, annotated with frequency counts, show shifts in subject popularity. Human intervention is crucial in deciphering the charts, to easily identify and compare the frequency of different subjects.

Task 2

Computation

The top 10 universities from 2011 to 2023 will be selected using the ‘groupby’ and ‘apply’ functions, and a line graph will be generated. To distinguish each university, distinct colours from 'tab20' colourmap will be utilised and user interaction is necessary here to know what colour map would be ideal and following the flow of Keim’s diagram, the visualisation should be refined to suit.

Visual

By representing each university with a unique colour, the viewers can easily track the performance of individual institutions to identify trends or patterns in their rankings. Human intervention is required as the visualisation could have displayed all 200 universities overwhelmingly and computational methods were necessary to group the ranks of each university across the years.

Task 3

Computation

Since the correlation matrix is calculated with the numeric data type, the rank-related features will be removed, and non-numeric features will be dropped. Human judgment is required as the features consist of other characters which converted to the object data type. Furthermore, human judgment helps identify the rank that belongs to the categorical data type and the desired features to be analysed.

Visual

The correlation matrix heatmap will be used to display the correlation between predictors and response. The colour intensity represents the strength and direction of correlation. The scatter plots with the regression lines will have the positioning and direction of the dots to indicate the strength of the correlation. Human judgment is required to investigate if multicollinearity exists, and the certain features need to be removed for constructing models.

Task 4

Computation

The dataset will be split into an 80% training set and a 20% test set. A multiple linear regression model will be used to predict the overall scores (response) with respect to score-based predictors. Also, the random forest will be used to predict the overall scores. To evaluate the model performance, the coefficient of determination will be calculated and compared between models.

Visual

The residual versus predicted value plot will be used to evaluate if the assumption of the linear relationship holds true. Also, the quantile-quantile plot (Q-Q plot) will be used to display the distribution of residuals following a normal distribution. Human judgment is needed to investigate if the model captures the underlying relationships in the data.

4.1 PROCESS

This section of the analysis is dedicated to answering the research questions.

Task 1

The countries are grouped into regions. For instance, Hong Kong and Korea are grouped as Asia, Canada and the United States are grouped as North America, and Ireland and the United Kingdom are grouped as Europe. Figure 4 shows the pie chart of the number of universities located in the region in 2011 and 2023. In 2011, Europe had the most top 200 universities with 51.9%, followed by North America with 23.7%. In 2023,

the number of top 200 universities in Europe dropped slightly to 47.5% while the number of top 200 universities in North America rose almost 50% to 32.5%. It is clear that over 12 years, Europe and North America still have the most top 200 universities.

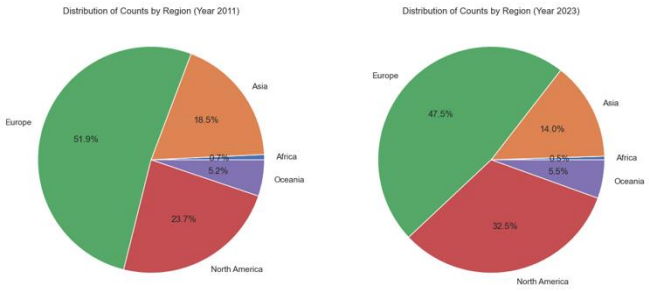

Figure 4: Pie charts showing distribution of universities across regions in years 2011 and 2023

To take the analysis further, a scoring distribution lens was added to the regional analysis and below the box plot highlights how North America generally exceeds the scoring distribution in 2011; however, in 2023, all other regions, where the top 200 universities have been grouped into, perform at very similar ranges. This observation could imply a more balanced and competitive landscape across various global regions over time. These charts used the same data transformations, so the comparative analysis was done on the same data frame ensuring consistency in analysis. The error bars in the 2023 chart are visibly larger which insinuates either a broader range of performance compared to the more tightly clustered scores observed in 2011 or it could reflect a higher degree of competitiveness among institutions [10].

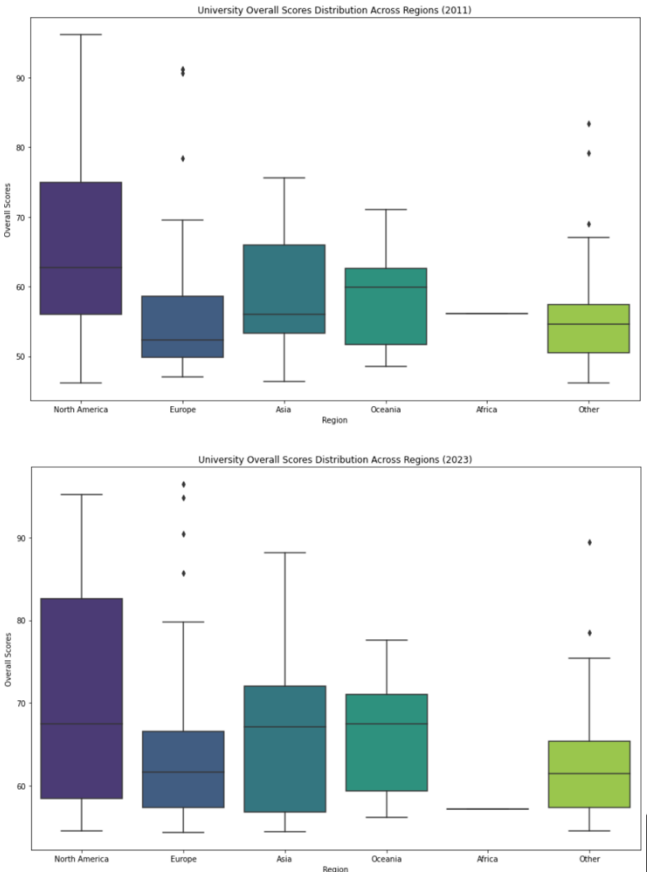

Figure 5: Box plots showing distribution of overall scores across regions in year 2011 and 2023

Figure 6 offers a very high-level perspective of comparing subject popularity in the years 2011 and 2023. At surface-level, there is definite consistency in the top 10 most popular subjects in these two years with ‘Biological Sciences’ and ‘Maths and Statistics’ leading the charts being taught in 194+ universities. There does seem to be a strong preference of STEM-based subjects across universities in both years whereas other disciplines only have single subjects taking a top position. Although, STEM subjects like ‘Electronic, Electrical and Mechanical Engineering’ are placed at the top of the 10 least popular subjects yet the annotation shows that it is still offered in 150+ universities therefore, this visualisation does lack context and should be treated as a snapshot of subject popularity distribution. It is worth noting that ‘Earth and Marine Sciences’ and ‘Environmental’ are both not on the charts in 2023 and subjects such as ‘Agriculture and Forestry’ and ‘Archeology’ are considered the 10 least popular subjects in both 2011 and 2023. We could infer that perhaps over

the years, this subject could have received lower funding meaning that it had to be removed from the academic offerings. [11]

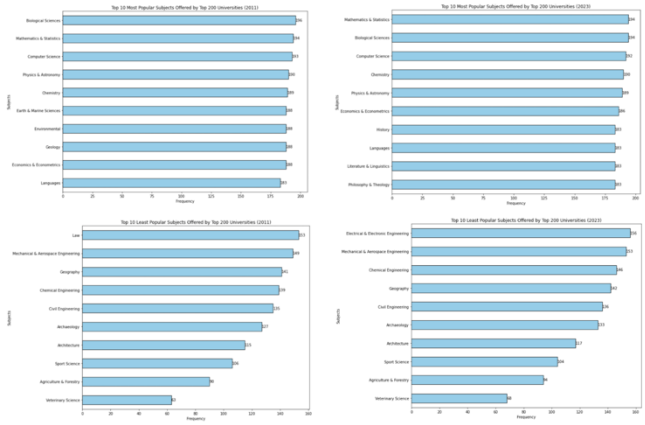

Figure 6: Bar charts showing both most popular and least popular subjects offered at universities in years 2011 and 2023

Task 2

Figure 7 depicts the temporal evolution of ranks of the top 10 universities from 2011 and 2023 and although the lines for each universities seem erratic there is consistency with regard to the universities listed on the charts. Only ETH Zurich seems to briefly appear on the chart from 2016 to 2019 whereas all other universities remain in the top 10 ranks across the years. Computer operations was essential to group the university's rankings across the years to develop this visualisation but without human intervention the observation regarding the temporary inclusion of ETH Zurich and the overall stability of other universities in the top 10 ranks would have been overlooked. Moreover, this visualisation steers away from potentially overwhelming displays of all 200 universities to focus on the more relevant top 10. The use of distinct colours and legends, not only makes this complex dataset more interpretable but also clearly highlights the evolution of each university. For example, Harvard University's trajectory fluctuates significantly compared to other universities jumping from ranks 1 in 2011 to 2 to 4 back to 2 then to 6 within 2 years. In 2020 its rank is at 7 but jumps back to Rank 2 in two years. An analyst would be able to instantly spot this erratic behaviour as the time series chart emphasises the volatility. Deciding to only show the top 10 universities became essential to highlight that there are still fluctuations and hidden behaviours even within the elite group of institutions that persistently rank and score high.

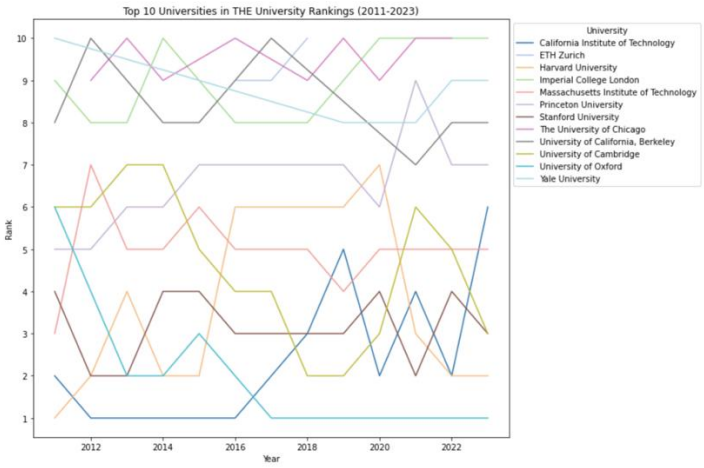

Figure 7: Temporal Analysis on top 10 universities' rankings from 2011 to 2023

Task 3

Non-numeric features, such as the location of the universities, and subjects offered, are removed. Also, the rank-related features, such as the overall rank, the rank of teaching scores, and the rank of research scores, are dropped as they are categorical data. Then the correlation matrix, based on the overall scores, teaching scores, international outlook scores, industry income scores, research scores, and citation scores, is calculated.

A correlation matrix heatmap (Figure 8) is displayed to show the strength of the correlation between different scores. Figure 8 illustrates that the research scores (predictor) have the strongest correlation with the overall scores (response) with a correlation coefficient of 0.91, followed by the teaching scores (predictor) with a correlation coefficient of 0.89. On the other hand, the international outlook scores and the industry income scores have a very weak correlation with the overall scores, with a correlation coefficient of 0.19 and 0.21 respectively. Also, it was discovered that the teaching scores have a very strong correlation with the research scores. This is reasonable as the professors and lecturers in the universities possess fruitful knowledge in which they can empower students in academic aspects and provoke students with critical thinking. To

avoid multicollinearity, when constructing the models, it is decided to remove the feature of teaching scores as it has a lower correlation coefficient with the overall scores. This is vital in building prediction models as the highly correlated features carry the same information which results in the increase of model complexity.

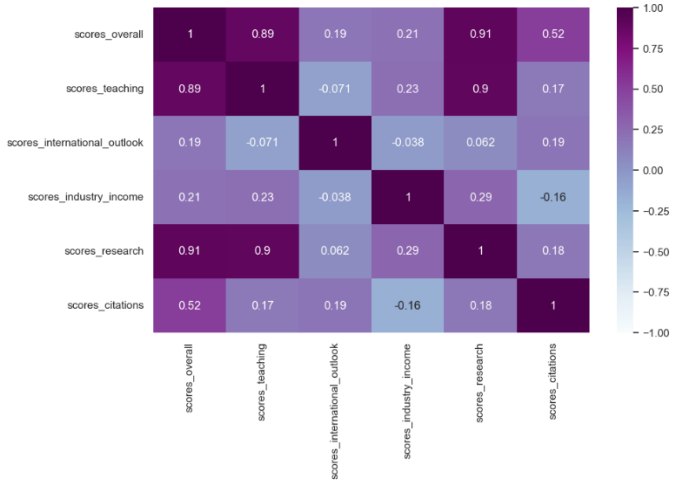

Figure 8: Correlation Matrix of features within dataset

The scatter plot is an alternative method to visualise the correlation between variables. Instead of color intensity, the scatter plot visualises the observations in a two-dimensional coordinate, providing a clear picture of the correlation. Figure 9 depicts the correlation between the overall scores (response) and the other scores (predictors). The scatter plot further verifies the results of the correlation matrix heatmap. It is clear that the dots in the scatter plot of overall scores and teaching scores, and the scatter plot of overall scores and research scores are more concentrated and closer to the regression line. On the other hand, the points in the scatter plot of overall scores and international outlook scores, and the plot of overall scores and industry income scores are more dispersed.

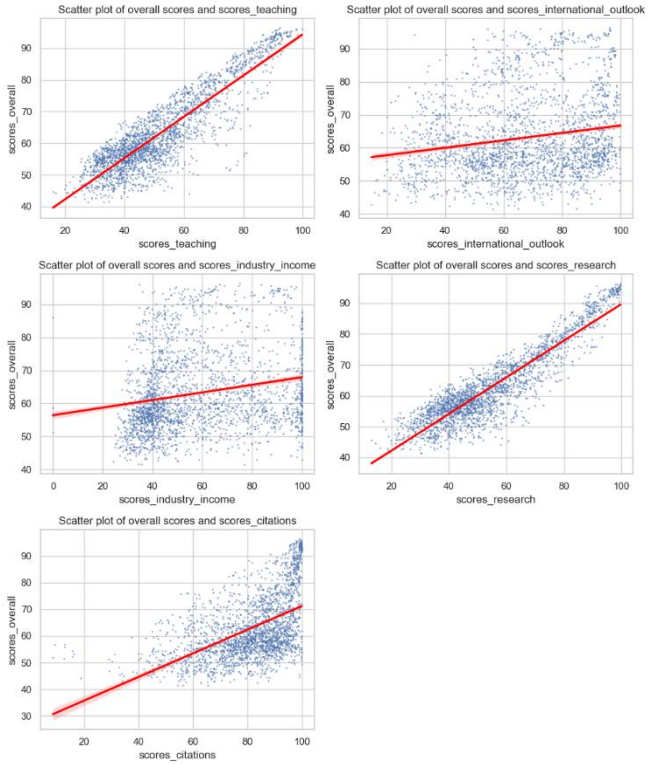

Figure 9: Scatter plots to show correlation of features

Task 4

From task 3, due to multicollinearity, the feature of teaching scores was removed. The four features, including the international outlook scores, industry income scores, research scores, and citation scores, were used to build the multiple linear regression model and random forest model for predicting the overall scores. The dataset was first divided into an 80% training set and a 20% test set. The models were then trained with the training set. To evaluate the performance of the models, the test set was used, and the coefficient of determination (R-squared) was calculated. The multiple linear regression model performs extremely well as the R squared is 0.9652. This represents 96.52% of the variance in the overall scores can be explained by the independent variables which are the international outlook scores, industry income scores, research scores, and citation scores. On the other hand, the random forest model performs even slightly better as the R squared is 0.9724. This indicates that 97.24% of the variance in the overall scores can be explained by the predictors.

The residual versus predicted value plot is a critical graph to evaluate the performance of

multiple linear regression models. Figure 10 shows that the random scatter of observations is located close to the horizontal line at zero. This indicates the assumption that the linear relationship holds true. It also shows that the spread of residuals is approximately constant in which the assumption of each set of values of the predictors having equal variance holds true.

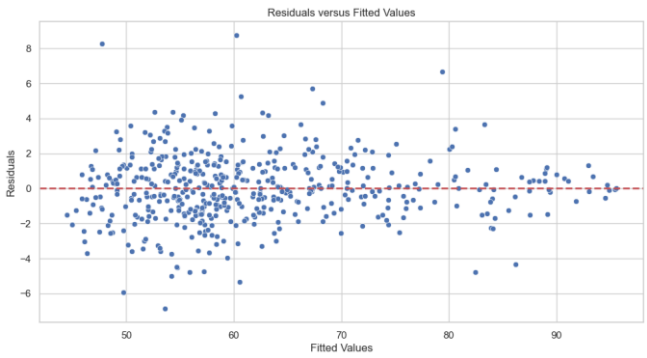

Figure 10: Scatter plot showing residuals vs fitted values

The quantile-quantile plot (QQ plot) is another powerful visualization for assessing if the dataset follows a specific distribution. Figure 11 shows that most of the points fall approximately along the diagonal line, despite the fact that a few points in the beginning and in the end deviate from the diagonal line, it is concluded that the residuals exhibit the normal distribution.

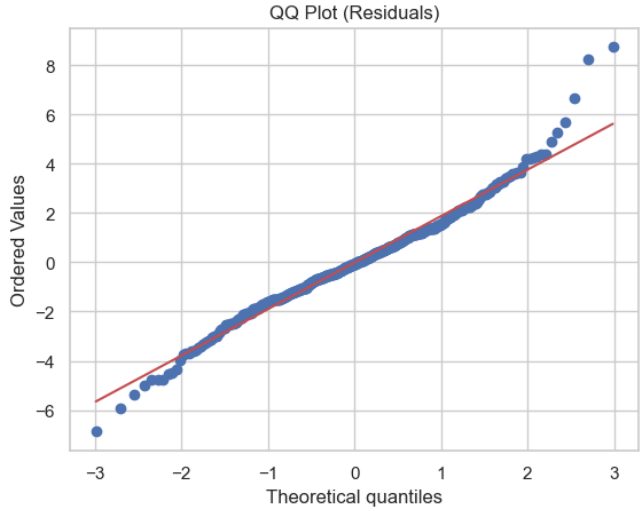

Figure 11: QQ Plot of residuals

Feature importance provides insights into which features contribute the most to the random forest performance. Figure 12 shows that the research scores contribute the most with the feature importance value of over 0.8 while the international outlook scores and the industry income scores contribute the least with the feature importance value of close to 0.

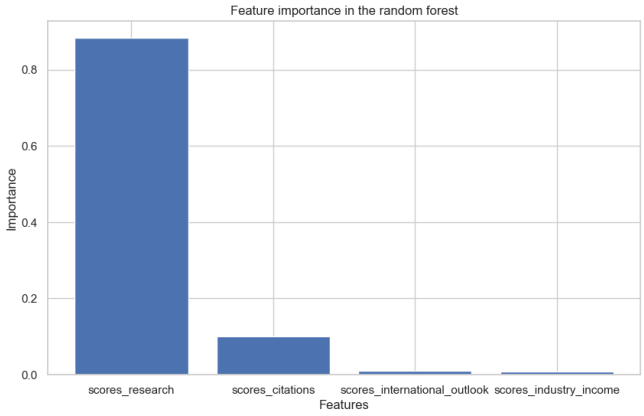

Figure 12: Bar chart showing feature importance

4.3 RESULTS

From the analysis, it became evident that not only were a significant proportion of universities located in North America, but North America also consistently held top rankings in the charts highlighted by figures 4 and 7 respectively. To answer question 1 the pie chart highlights its regional prominence, and the time series analysis further proves how universities within North America consistently dominated the top 10 positions throughout the years, addressing question 2. It is interesting to note, however, that the box plot in figure 5 in fact reveals that other regions scored similarly to North America in 2023 and yet could not dominate the top 10 ranks.

Moreover, since our research domain sits within Education, it was interesting to find that STEM subjects dominated the most popular subjects offered across 200 universities and the correlation matrix heatmap revealed that ‘teaching_scores’ had the highest influence on the overall scoring. The scatter plot provided an alternative method to visualise the correlation, validating the profound correlation between the overall scores, the teaching scores, and the research scores. Once the predictive models were built, it was pleasing to observe that the multiple linear regression models and the random forest models both performed well, with the latter model performing better. The residual versus predicted

values plot and Q-Q plot verified the assumption of linearity and normality respectively for the multiple linear regression model, while the feature importance chart depicted that the research scores are the most influential features in the random forest model.

4.3 CRITICAL REFLECTION

When approaching the analysis to demonstrate the geographical, scoring and subject preference distribution, the decision to choose 2011 and 2023 as the years was to potentially emphasise a significant difference however, this assumption may not be the most robust way of comparative analysis. Yet, to ensure that no generalised conclusions were made, analysing several slices of the data allowed for a well-informed interpretation of the dataset. Thinking strategically, it would have been ideal to do the State of the Art analysis much earlier in the research planning stage as it would have informed the visualisation approach more and reinforced confidence sooner that the charts do in fact answer the research question quite directly. However, there were limitations to the data when the temporal analysis was conducted as there was limited contextual information about potential external factors affecting the fluctuations such as economic conditions [12] and student sentiment, etc.

The horizontal bar chart used to show subject preference is a basic representation and may be seen as a weak technique to show distribution where perhaps a clustering algorithm could have been applied instead to group universities based on similar subject offerings. However, when developing the correlation matrix, the column 'subject_offerings' was removed as it was seen as categorical data whereas scoring is numeric therefore the relationship may not have actually been reliable. To avoid multicollinearity and model complexity, the strategic decision is to exclude the teaching scores when constructing the predictive models. The scatter plot between the overall scores and the underlying factors serves as a compelling visual validation, reinforcing the earlier findings with a representation of the distribution of data points. This analytical journey highlights the importance of feature selection and builds a solid foundation for constructing robust predictive models.

Under the approach of dividing the dataset into an 80% training set and a 20% test set, the deployment of multiple linear regression model and random forest model unveils excellent predictive power. The latter model provides a slightly better performance in terms of R squared to explain the variance of independent variables by the predictors. On the other hand, the analysis of multiple linear regression model assumptions through the residual versus predicted value plot and quantile-quantile plot verify the robustness of the linear relationship and the normality of residuals. Also, the analysis of the random forest model through feature importance provides insights that the research scores have a dominating feature importance value and contribute the most to the model performance. In summary, the analysis not only affirms the efficacy of the predictive models but also underscores the process of human reasoning required in computational and visual analysis.

Improved software functionality, perhaps an intuitive model tuning interface, could have enhanced the model as it allows analysts to easily experiment with hyperparameters in real-time [13]. This interactive feature would facilitate the exploration of various parameter combinations, enabling a more efficient identification of the optimal settings for the random forest model. Additionally, incorporating automated model performance metrics within the software could streamline the assessment process, providing instant feedback on different model configurations. This would save time and inform decisions, contributing to the overall improvement of the predictive modeling process. To anyone conducting this research in the future should consider applying robust imputation techniques, such as k-nearest neighbors [14] rather than removing the null values completely which could have in turn made the prediction models more accurate.

REFERENCES

[1] Hou, Ya-Wen & Jacob, W.J.. (2017). What contributes more to the ranking of higher education institutions? A comparison of three

world university rankings. International Education Journal. 16. 29-46. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Estrada-Real, A.C., Cantu-Ortiz, F.J. A data analytics approach for university competitiveness: the QS world university rankings. *Int J Interact Des Manuf* **16**, 871–891 (2022). https://doi.org/10.1007/s12008-022-00966-2

[3] Tabassum, Anika & Hasan, Mahamudul & Ahmed, Shibbir & Tasmin, Rahnuma & Abdullah, Deen & Musharrat, Tasnim. (2017). University ranking prediction system by analyzing influential global performance indicators. 126-131. 10.1109/KST.2017.7886119.

[4] Bothwell, E., Jack, P., Ross, J., Thomas-Alexander, T., Philpotts, E., Gill, J., Ellis, R., Kyte, R., Torrealba, C., Stott, E., Day, C., digital, T., & Lem, P. (2023, November 30). *World University Rankings*. Times Higher Education (THE). https://www.timeshighereducation.com/world-university-rankings#:~:text=THE%20(Times%20Higher%20Education)%20has,governments%20and%20industry%2C%20since%202004.

[5] Kapatsyn, A. (2023, April 3). The World University Rankings 2011-2023. Kaggle. https://www.kaggle.com/datasets/r1chardson/the-world-university-rankings-2011-2023/data

[6] Walpole, R.E., Myers, R.H., Myers, S.L. and Ye, K. (2007) Probability & Statistics for Engineers & Scientists. 9th Edition, Pearson Education, Inc.

[7] Visual Analytics Towards Tool Interoperabilty - A Position Paper - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Visual-analytics-process-defined-by-Keim-et-al-Keim-et-al-2008_fig1_301721746

[8] Harrison, J. (2015). *Regional Geography*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780080970868720435

[9] Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. Psychological Science in the Public Interest, 22(3), 110-161. https://doi.org/10.1177/15291006211051956

[10] Hart, P. and Rodgers, W. (2023). Competition, competitiveness, and competitive advantage in higher education institutions: a systematic literature review. *Studies in Higher Education*, pp.1–25. doi:https://doi.org/10.1080/03075079.2023.2293926.

[11] Geoscience on the chopping block. *Nat Rev Earth Environ* **2**, 587 (2021). https://doi.org/10.1038/s43017-021-00216-1

[12] Valero, A. and Van Reenen, J. (2018). The Economic Impact of Universities: Evidence from Across the Globe. *Economics of Education Review*, [online] 68. doi:https://doi.org/10.1016/j.econedurev.2018.09.001.

[13] Passerat-Palmbach, J., Matache, C. and Kainz, B. (2019). *MEng Individual Project Efficient Design of Machine Learning Hyperparameter Optimizers*. [online] Available at: https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1819-ug-projects/MatacheC-Efficient-Design-of-Machine-Learning-Hyperparameter-Optimizers.pdf [Accessed 7 Jan. 2024].

[14] Alam, S., Ayub, M.S., Arora, S. and Khan, M.A. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, [online] 9, p.100341. doi:https://doi.org/10.1016/j.dajour.2023.100341.

Table of word counts (Pair work)

Problem statement	252
State of the art	452
Properties of the data	506
Analysis: Approach	619
Analysis: Process	1544
Analysis: Results	248
Critical reflection	563