ASSESSMENT SUBMISSION

Module Title:	Principles of Data Science
Module Code:	KL7010
Academic Year / Semester:	2023-24 / Semester 1
Module Tutor / Email (all queries):	David Hastings david2.hastings@northumbria.ac.uk
% Weighting (to overall module):	75%
Assessment Title:	Assignment 1 - The selection, application and evaluation of data science methods, tools and techniques.
Date of Handout to Students:	Week commencing 2nd October 2023
Mechanism for Handout:	Module Blackboard Site
Deadline for Submission Attempt by Students:	15th January 2024 on or before 11.59pm GMT
Mechanism for Submission:	Document upload to Module Blackboard Site
Submission Format / Word Count	Please upload your written report as a single PDF document. Your report should not exceed 3,000 words in length, not including the front cover, contents page, references, bibliography and appendices
Date by which Work, Feedback and Marks will be returned:	9 th February 2024
Mechanism for return of Feedback and Marks:	Mark and individual written feedback will be uploaded to the Module Site on Blackboard. For further queries please email module tutor.
Student ID	W23023023
Student Name	Arka mandol
Word Count	2954

The Selection, Application and Evaluation of Data Science Methods, Tools and Techniques.

INTRODUCTION

The ranking of universities is a complicated and multi factored process. It has a huge impact on the students, academic institutions, governments, and industries too. Different methods and criterias are involved in global university ranking. Majority of the students check for the ranking when it comes to taking admission in a university. Universities assess the ranking to make themself better in the fields they lack (Tabassum et al., 2017, 126).
Generally all the techniques and core concepts of ML(machine learning) & data analysis are too much helpful to explain the past and also for predicting the future (Han et al., 2022,) and further data exploration (Power et al., 2015).
Different types of methods are used by the national and international university ranking systems. In the United Kingdom we have The Times Higher Education World University ranking and this ranking is followed by all the universities and is a very influential ranking system. SImilarly there are many other university rank providers like CWUR from Saudi Arabia, Shanghai ranking from china,etc. The Times Higher Education World University Ranking provides us with the global ranking of universities basically on the basis of their performance in research, teaching, international students, etc. The ranking is generally done on basis of 13 parameters that provide unbiased comparisons that are trusted by the students, professors, university staff, organisations and government too (Tabassum et al., 2017, 126).
This report contains the overall analysis of the World University Ranking dataset. In field of data science the processes involved are as important as the data and It is important to have a good plan , continuous knowledge about the tools and methods in the fields of data science. This is important for our task of predicting University ranking based on the different parameters provided.
The main components of the analysis are to explore the data, select the best analysis method, preprocessing of the dataset, building the model, evaluating the model, and the presentation of the finding throughout the process and these report faithfully follows the fundamentals of data science and able to offer finding out the areas which can help in improvement in university rankings.

METHODOLOGIES

The process involved for data analysis for predicting the overall score for the university ranking are as follows:-

1. Data Exploration and Understanding:

The main purpose of this process is to get a better understanding and familiarise with the dataset's structure, characteristics and limitations. Which are important to carry out next stages (Provost & Fawcett, 2013). In this part we use R packages functions like 'read_csv()' to load the dataset followed by 'summary()', 'head()', 'str()' functions to get the initial overview of the data. These initial checks help us in finding the type of the columns, finding the missing values, and help in understanding the dataset's layout. This is a crucial and important initial step for all data science tasks(Wickham and Grolemund 2017).

2. Data cleaning and Preprocessing:

After we are able to load the data successfully the next important phase is preprocessing and cleaning the data to continue for further processes and the importance of this process part cannot be overstated though the quality directly impacts the reliability of the analysis(Larose & Larose, 2014).

This part generally involves:

Handling the Non-Numeric Values:

In our university ranking dataset the fields like ' stats_number_students' and 'stats_pc_intl_students' have numeric values but not in standard form (for example, numbers with commas and percentages sign)and that's why it is important to convert these fields to numeric format for quantitative analysis.

Dealing with Missing Data:

Missing value can have a huge impact on the quality of the prediction and we need to handle the missingness carefully. Important decisions need to be taken in this part like imputing, removing record or replacing the na values in this part after doing a thorough study of missingness, whether the missingness are Missing Completely at Random (MCAR), Missing At Random (MAR) or Missing Not at Random (MNAR).

Handling missing data with the process of imputation is preferred over outright deletion, mostly in the datasets where omissions can lead to crucial information loss(Hastie et al., 2013).

Data Type Conversion:

We need to make sure of that each variable has a correct data type and it is important. For example, factor values need to be converted to numeric , symbols need to be removed if present, etc.

Creating New Variables:

In some cases we need to derive new features from the existing features to capture more better relationships within the data. For example we could calculate the ratio of international students with respect to domestic students(not required in our scenario).

The step that were required for our university ranking datasets in this process are:-

Firstly, I created a function to convert the string type “n/a” field to Null(not available value, so that i can use inbuilt r functions to find the na values in our dataset). The missingness is around **28.769 % ,** meaning rows which contain at least one na value in it (Did Not considered ‘stats_male_female_ratio’ as it was handled and ‘subject offered’ columns because it was removed).

Secondly, I removed the constant value column (‘closed’ field) and column having textual values, which are not required for further EDA like the ‘subjects_offered’.

Thirdly, I needed to make an amputation of the NA fields as the data is MNAR (Not missing at random), as we can see in the below in Fig.1.

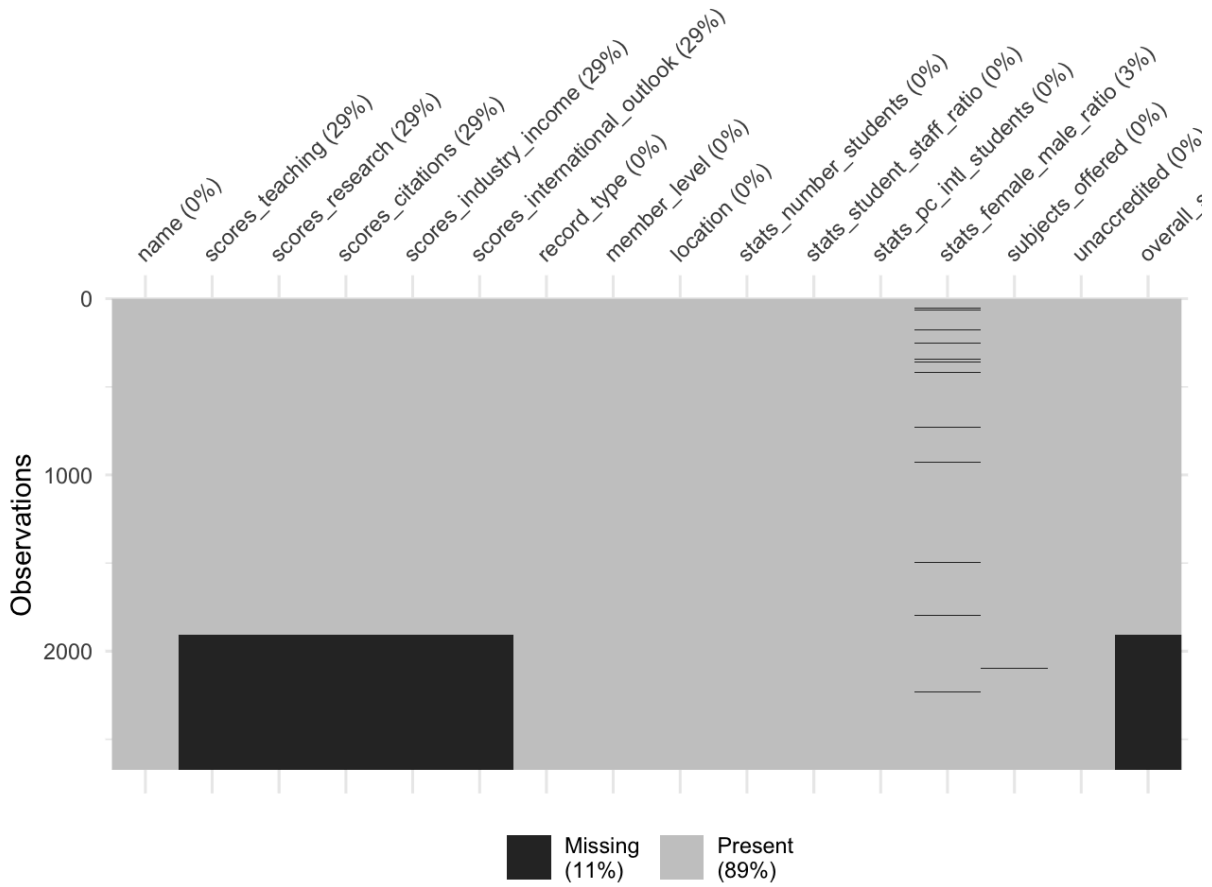


Fig.1. Representing the missingness of values in the whole dataset.

Fourthly, I created a function to replace the values which were in range in the ‘overall_score’ column with the avg values. Then, Converted 'stats_number_students' to numeric by removing commas, Converted percentages in 'stats_pc_intl_students' to numeric, Ensured numeric data types for plotting

Fifthly, Handled 'stats_female_male_ratio' column by making it two columns(‘male_ratio’, ’female_ratio’) and the na values in the columns were replaced by mean of the columns.

3. Exploratory Data Analysis (EDA):

Exploratory data analysis is a critical part of the required for prediction of value from a dataset. It is the part where we understand and find the patterns, relationships and anomalies within the data(Tukey, 1977).

EDA is important for gathering deep information from the dataset. This step generally involves:

Statistical Summaries:

Here we examine the and check for the central tendencies, distributions and dispersions if present in the dataset. The basic behaviour can be understood here from the data.

Visualisations:

Here we plot all the graphical representations of the data. We basically view the histograms, box plots, and scatter plots for visualising distributions , relationships and patterns in the dataset, which helps a lot going further for feature selection. In the R language we use the ‘ggplot2’ library, which helps in providing a wide range of informative graphs.

Correlation Analysis:

In this part we understand how every feature is correlated with each other and the relationship of all the features with each other and more importantly to the target variable that needs to be predicted(‘overall_score’ in our university ranking scenario). It is basically one of the most crucial parts in the exploratory data analysis part. So generally, correlation coefficients and plots can determine the important features that need to be used in the next process.

Now for our university ranking dataset we ensured numeric data types for plotting, handaled NAs in these columns if they exist further.

Further plotted scatter plots of key features('scores_teaching', 'scores_research', 'scores_citations', 'scores_international_outlook', etc) against the overall score.
As shown in the below figures

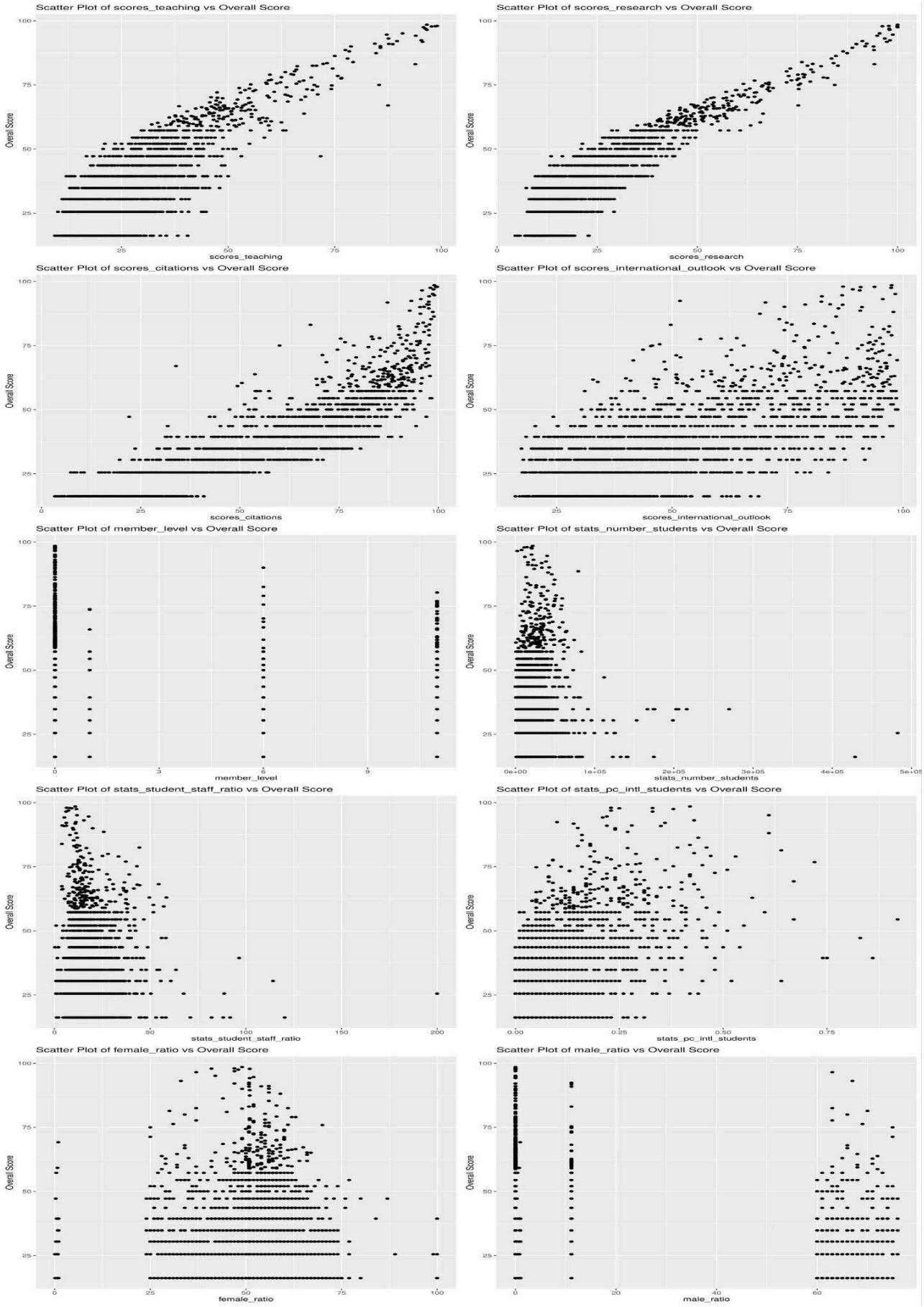


Fig.2 Representing all the scatter plots to show relation of feature with overall_score

After plotting the scatter plots, plotted the correlation matrix with the features after cleaning the data.

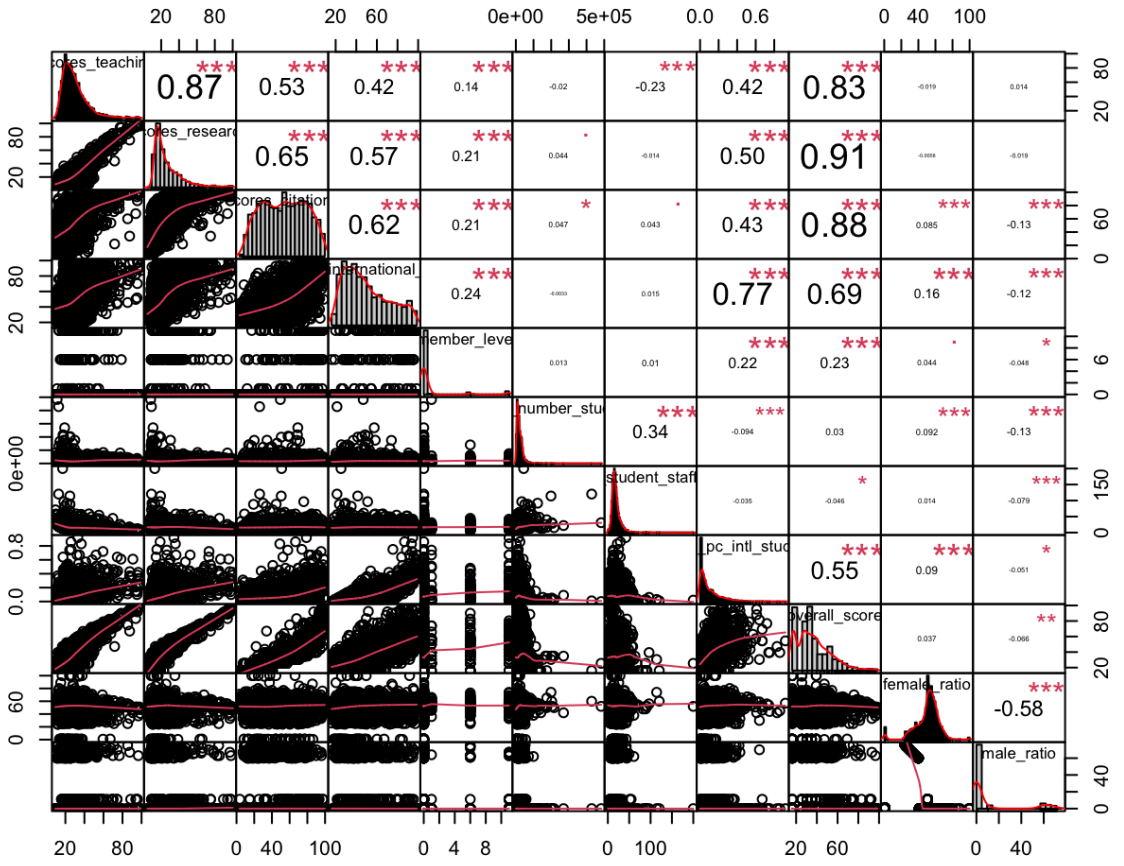


Fig.3 Representing the correlation graph along with Histogram

The exploratory data analysis (EDA) provides us with the following valuable insights:

1. Correlation Matrix:
   From Fig 3 and Fig 4, heat map and the correlation diagram it is observed that features like 'scores_teaching', 'scores_research', and 'scores_citations' have a high positive correlation with the 'overall_score', indicating they are significant predictors.
2. Scatter Plots:
   Similarly the scatter plot suggests that the higher the scores in teaching, research, and citations then higher overall scores.

So based on the analysis we can say that features like 'scores_teaching',

'scores_research', 'scores_citations', and 'scores_international_outlook' are likely to be the most important predictors for the overall score.

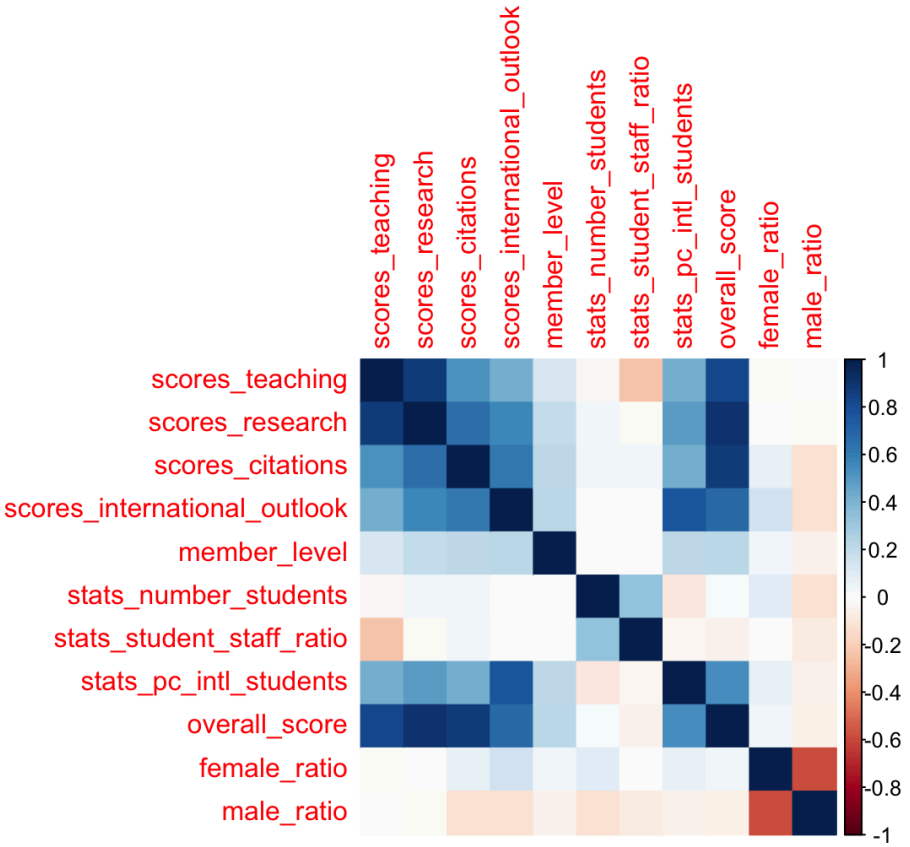


Fig.4 Representing the heat map of the correlations

4. Feature Selection & Model Building:

Feature selection & model building are the core part of predictive analytics. It is the part where we build the required machine learning model to do the final prediction. We need to follow few steps before modelling such as:

Feature selection:

So, after looking at Exploratory Data Analysis, we can decide and select the models that are required for the predictive model.

Data splitting:

Further we need to split the data to train and testing sets. The train set will be used to train selected models and the testing set will be used later to evaluate the model on unseen data.

Normalisation/Standardization:

The data need to be normalised or standardised based on our predictive model, so that the data is suitable to use for the predictive model.

For our university ranking dataset the selection of the feature is directly related to the model prediction performance and interpretability. Analysing the university ranking data we can come to the conclusion that the features like teaching quality, research output, and citations are the most important features based on the correlation with the target field overall score.

Looking into the nature of our target field overall_score we can say that it is a not a classification problem rather it is a regression problem because :

1.Continuous Target Variable: The overall_score is a continuous variable and that makes it very ideal for regression analysis.
2.Predicting Score: Regression models(We use when we want to know how much or how many and want num value returns) are used and designed for predicting numerical values based on the input features given (Bonaccorso, 2017).

I have selected regression models like Linear Regression, Decision Trees, and Random Forest as these models suit perfectly with the nature of the problem and the characteristics of the dataset. For instance Random Forest is a suitable model that tends to avoid overfitting and also can handle non-linear relationships perfectly in most cases(Breiman, 2001).

The training and testing of models is done post splitting the data, this is generally a standard practice in data science, which allows the assessment of a model's performance on unseen data and makes sure of its generalizability (James et al., 2021). For our university ranking dataset we have used the ‘createDataPartition()’ function to split the data and the ‘p’ value is set to 0.8 (p = .8).

In order to implement the ML models in R I used libraries ‘caret’ for model training and ‘randomForest’ for the random forest algorithm. These libraries are well equipped and provide a framework to work for machine learning tasks, simplifying the steps to build the model and evaluating it.

5. Model Evaluation and Practical Appraisal

Performance Metrics:

In a dataset like university ranking, where we have to predict the overall score by regression model are generally evaluated using metrics like Mean Squared Error (MSE) and R-squared (R^2). Mean Squared Error gives us the measure of error for the model prediction and R^2 gives us the proportion of variance in the dependent variable explained by the model(Montgomery et al., 2012). These metrics basically offer us the quantitative assessment of the models' predictive capabilities.

Feature Importance Analysis:

Specially for using the model Random Forest we conduct an analysis of the features that are important. This Random Forest analysis gives us the features that are important and has a huge impact on the predictions, offering insights into the underlying dataset (Strobl et al., 2007).

Critical Evaluation of Model Performance:

Apart from Quantitative metrics , we have more crucial assessments of the models like complexity of the model, model interpretability, and also the nature of the data. For Example we know that Random Forest might provide high accuracy but it is less interpretable then LM model (Linear regression model) and these factors might need to be crucial dependent on the application(Louppe, 2014).

Cross-Validation:

In order to check if the model is overfitting or not and to assess their generalizability, we used cross-validation techniques. Generally in the cross validation process we divide the dataset in multiple subsets and then we validate the model for all the subsets. This technique provides a more robust evaluation of the model's performance across different data samples (Kohavi, 1995).

Comparative Analysis:

In this last step we compare all the models that were trained and evaluated and get to know the performance of the models against each other. This helps us to get a practical appraisal of their relative effectiveness. Most importantly the comparison basically helps us to identify the best suitable model for the required task.

We will be not using confusion matrix in this issue, because confusion matrix is generally used for classification issues not in regression issues. In our regression problem our goal is to predict a continuous output. To check the performance here is measured by knowing how close the predicted values are close to the actual values. The commonly used metrics for evaluating the regression models are . In Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. In our scenario we will be using the MSE and R-squared values for evaluating the models.

After successfully training and evaluating the models for our University ranking datasets. Below are the results:

For Linear Regression:

- The Mean Squared Error (MSE) is: 4.82
- The R^2 Score is: 0.98

For Decision Tree:

- The Mean Squared Error (MSE) is: 25.69
- The R^2 Score is: 0.91

For Random Forest:

- The Mean Squared Error (MSE) is: 4.04
- The R^2 Score is: 0.98

By the above results we can come to the confirm the following :

- For the University Ranking dataset, we can consider Random Forest is the best predictive model compared to the other three models because it has the lowest MSE and the highest R^2 score.
- Also the LM (Linear Regression) model performs good for the university ranking dataset, it just has Slightly higher MSE and slightly lower R^2 score then Random forest model .

- Coming to the third model, Decision Tree has Higher MSE(Mean Squared Error) and less R^2 score making it not suitable for our University ranking dataset.

Now based on the results it can be said that Random Forest is the best suited choice for predicting the overall_score of the universities in the university ranking dataset.

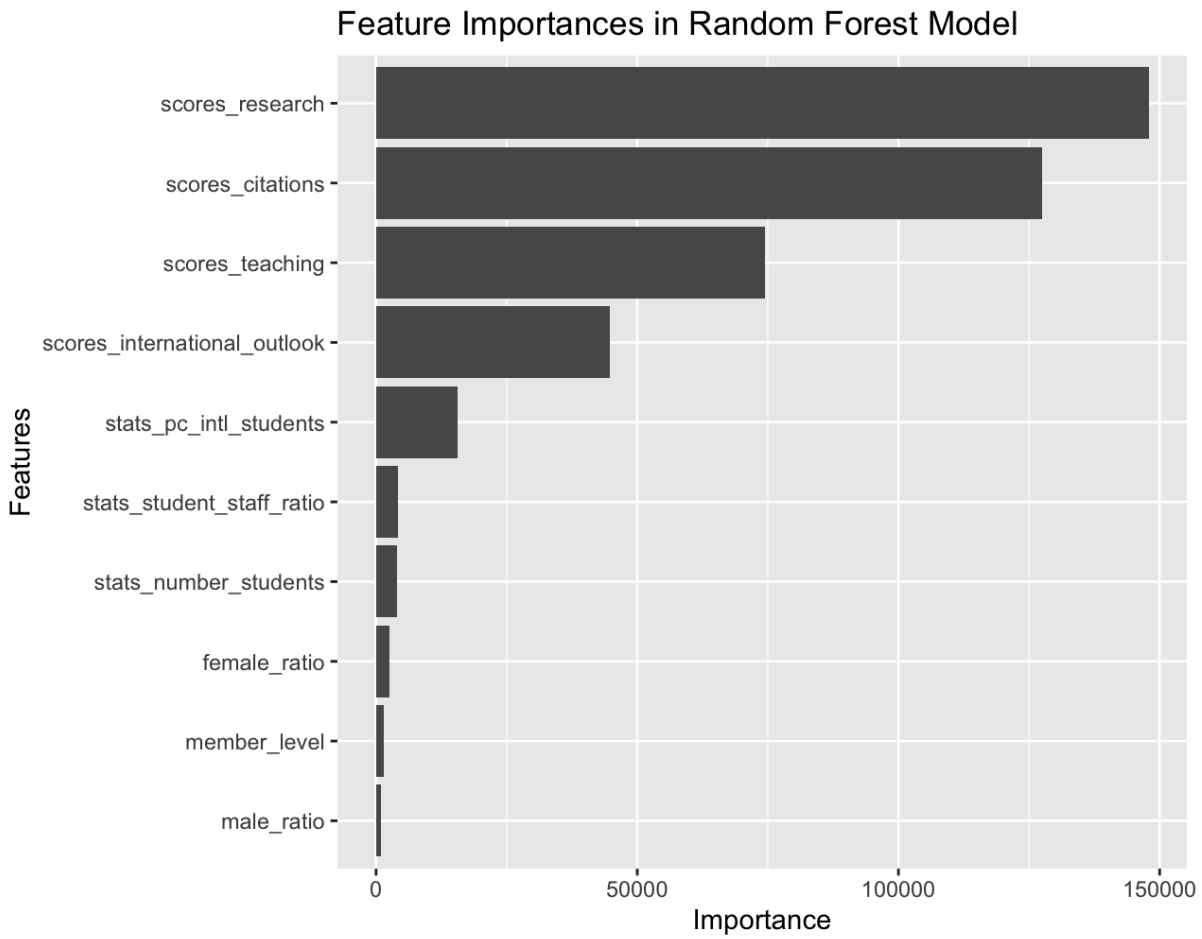
6. Feature Importance:

The feature importance of the Random Forest model give us the below outputs:

Scores Research: (importance 34.93%)
This is the most influential feature in predicting the overall score for our university ranking dataset. This indicates that if the research score of the university is good the overall rank is also improved.
Scores Citations: (importance 30.10%)
This is the second most important feature of the university ranking dataset and tells us that Citations are also the key factor in recognition of the university's research in the academic community and plays a major role in overall_score.
Scores Teaching: (importance 17.60%)
Although it is an important part of universities, our model tells us that teaching score has less effect on the overall score as compared to Research and citations.
Scores International Outlook: (importance 10.57%):
Compared with the above factors this feature has less influence on the overall score of the university.

Rest features like Stats Percentage of International Students (importance 3.70%), Stats Number of Students(importance 0.95%), Stats Student-Staff Ratio (importance 0.99%), Female Ratio (importance 0.61%), Member Level (importance 0.35%), Male Ratio (importance 0.20%) has very little influence to the overall score according the model and university ranking dataset.
Below diagram represents the feature importance in graphical representation.

These results tell us the importance of research quality and citations in stating the University overall ranking. Also tell that teaching quality and international outlook are important too but not as decisive as research performance.



Conclusion

In conclusion, the analysis of the World University Ranking dataset by using ML (Machine Learning) models in R language provided the required insightful outcomes for predicting the University Rankings. The methodologies consist of data exploration, cleaning, and preprocessing, continued by a thorough EDA (exploratory data analysis) for finding out the key predictors for the overall ranking. This report found that features like 'scores_teaching', 'scores_research', 'scores_citations', and 'scores_international_outlook' are the most influential features in predicting the university's overall score.

Among the three models(Linear regression, Decision tree, Random forest) used for the report, the Random Forest model has come out to be most effective. The Random forest model gave the lowest Mean Squared Error(4.04) and the highest R^2 score(0.98).The superiority highlights the importance of balancing accuracy and interpretability in our predictive models.

Further in the feature important analysis part shows us the importance of research quality and citations in determining a university's ranking. Also we saw that teaching quality and international outlook has less significance but also has an impact on the overall university ranking.

Overall, In this report we can see that the machine learning techniques provide us a robust framework for better understanding and predicting universities ranking and do provide valuable insights for the academic institutions and stakeholders in the educational sector.

References

Bonaccorso, G. (2017). *Machine Learning Algorithms*. Packt Publishing. https://books.google.co.uk/books?hl=en&lr=&id=_-ZDDwAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning+algorithms&ots=eqdDz4ED2D&sig=-vVZnV2vEs-4jvyv15_5rb-S4Qw&redir_esc=y#v=onepage&q=machine%20learning%20algorithms&f=false

Breiman, L. (2001, October). Machine Learning. *Random Forests*, *45*. https://doi.org/10.1023/A:1010933404324

Han, J., Pei, J., & Tong, H. (2022). *Data Mining: Concepts and Techniques*. Elsevier Science. https://books.google.co.uk/books?hl=en&lr=&id=NR1oEAAAQBAJ&oi=fnd&pg=PP1&dq=J.+Han,+J.+Pei,+and+M.+Kamber,+Data+mining:+concepts+and+techniques.+Elsevier,+2011&ots=_N1HMNtfo_&sig=7NalSBVC-ofhD_8msdp8oMDl7IA#v=onepage&q&f=false

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (1st ed.). Springer New York. https://doi.org/10.1007/978-0-387-21606-5

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137-1145.

https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley. Retrieved January 01, 2024, from https://books.google.co.uk/books?hl=en&lr=&id=9hOpAwAAQBAJ&oi=fnd&pg=PR11&dq=(Larose+%26+Larose,+2014)&ots=9R2y7VdMQ7&sig=LhsdT75gaDs4dmHJaOniNUxhJlE#v=onepage&q=(Larose%20%26%20Larose%2C%202014)&f=false

Louppe, G. (2014). Understanding Random Forests: From Theory to Practice. *arXiv preprint arXiv:1407.7502*. https://doi.org/10.48550/arXiv.1407.7502

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley. https://ocd.lcwu.edu.pk/cfiles/Statistics/Stat-503/IntroductiontoLinearRegressionAnalysisbyDouglasC.MontgomeryElizabethA.PeckG.GeoffreyViningz-lib.org.pdf

Power, D. j., Sharda, R., & Burstein, F. (2015). *Decision Support Systems* (C. L. Cooper, Ed.; Vol. 7). Wiley. https://doi.org/10.1002/9781118785317.weom070211

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51-59. https://doi.org/10.1089/big.2013.1508

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(25). https://doi.org/10.1186/1471-2105-8-25

Tabassum, A., Hasan, M., Ahmed, S., Tasmin, R., Abdullah, D. M., & Musharrat, T. (2017). University ranking prediction system by analyzing influential global performance indicators. In *2017 9th International Conference on Knowledge and Smart Technology (KST)* (pp. 126-131). IEEE. 10.1109/KST.2017.7886119

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company. https://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf

Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly. Retrieved January 14, 2024, from https://r4ds.had.co.nz/