HDSC August ’22 Premiere Project Presentation

By

Team Deep Learning

Topic: University Rankings for Years 2018, 2019 and 2020

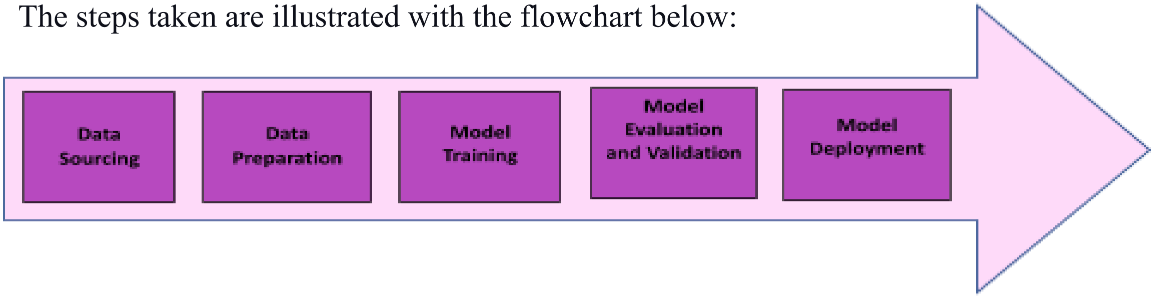
Every year, thousands of people make the decision to attend or build careers in a specific university. Choosing a university can be an overwhelming task and there are many factors to consider. The current university ranking system, such as the US News and World Report rankings, is a major factor in students’ decisions about where to attend school. Rankings are based on a variety of factors such as student-faculty ratios, graduation rates, and the amount and quality of the courses offered. University rankings also take into account institutional reputation and other factors that students value. While these rankings provide students with important information about the quality of a particular school, they should not be the only source of information used when making this important decision.

Aim and Objectives

The aim of this project is to determine the metrics that contribute to a university’s rank and to deploy a machine learning model that predicts the overall score of a university based on past university rankings and records. The project also intends to determine how certain Classifications such as (institution's size, subject range, research intensity, age, status) influence the model.

Flow Process

The steps taken are illustrated with the flowchart below:



Data Sourcing

Data Preparation

Model Training

Model Evaluation and Validation

Model Deployment

Data Sourcing

For the problem being addressed here, we used open-source data on Kaggle and can be accessed via the link below:

[https://www.kaggle.com/datasets/divyansh22/qs-world-university-rankings]

Data Preparation

The following steps were followed in the data preparation process:

- **Data collection:** The process of data preparation starts with gathering data containing certain elements which contribute to noise in the analysis and model building.

- **Data discovery and profiling:** The 3-year datasets contained lots of missing values and special characters such as '+', '-' and '=' which were observed and removed from the data. The columns were correctly renamed.
- **Data cleaning:** It involved cleaning special characters, symbols, and non-numeric attributes that were not needed from the dataset. Also, several transformations were done on the data, like converting columns with string data type to numeric data type to make them fit for analysis and modelling. The rows having missing values were removed from the dataset as well as data that does not contribute to the overall performance of the analysis and modelling.
- **Data structuring:** After the datasets have been satisfied as to be free from any irregularities, the cleaned datasets were stored in the comma-separated values file.
- **Data transformation:** The data transformation involves adding a feature that helps in understanding the behaviour of the dataset.
- **Data visualisation:** This last step was done to enable us to understand the prepared dataset through visualisations plots like histograms, scatter plots, box plots, line plots, and bar charts.

Exploratory Data Analysis

After the cleaning process was done, visualisation was done to ascertain the behaviour of the dataset. The correlation heatmap shown below indicates the relationship between the features. While a few variables show some degree of relationship (like academic_reputation_score and employer_reputation_score), it could be said that the dataset is largely not multicollinear. The bottom row shows that each of the variables is linearly related to the target variable (Overall_Score).

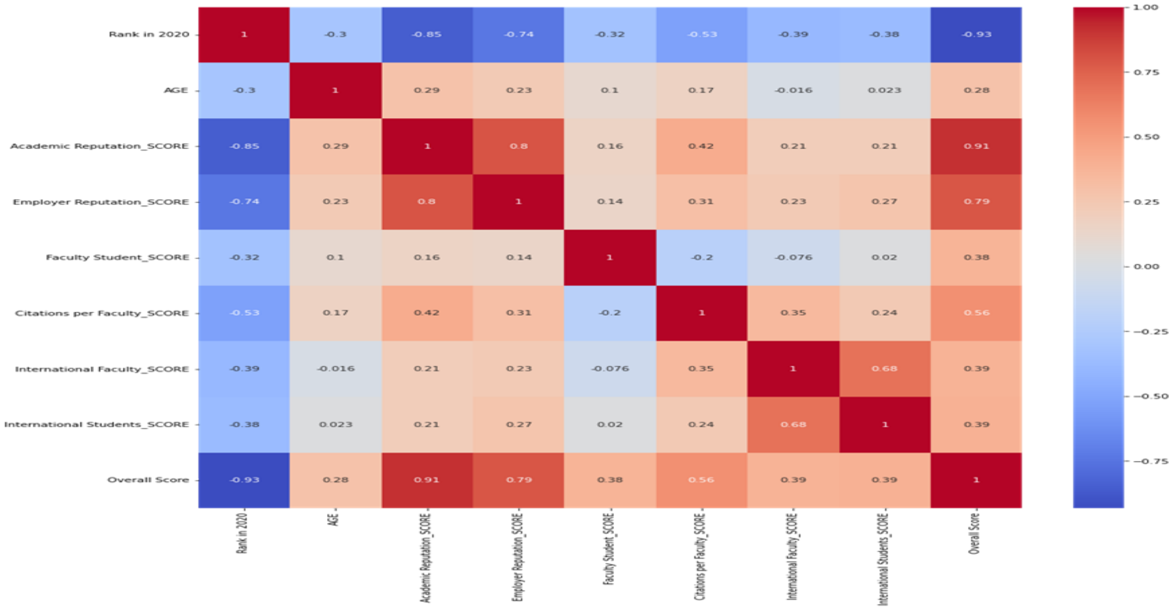


Figure 1: Correlation Heatmap

The box plot shown below was used to check for outliers in the cleaned data. Of course, there were no visible outliers as we can see in the box plot.

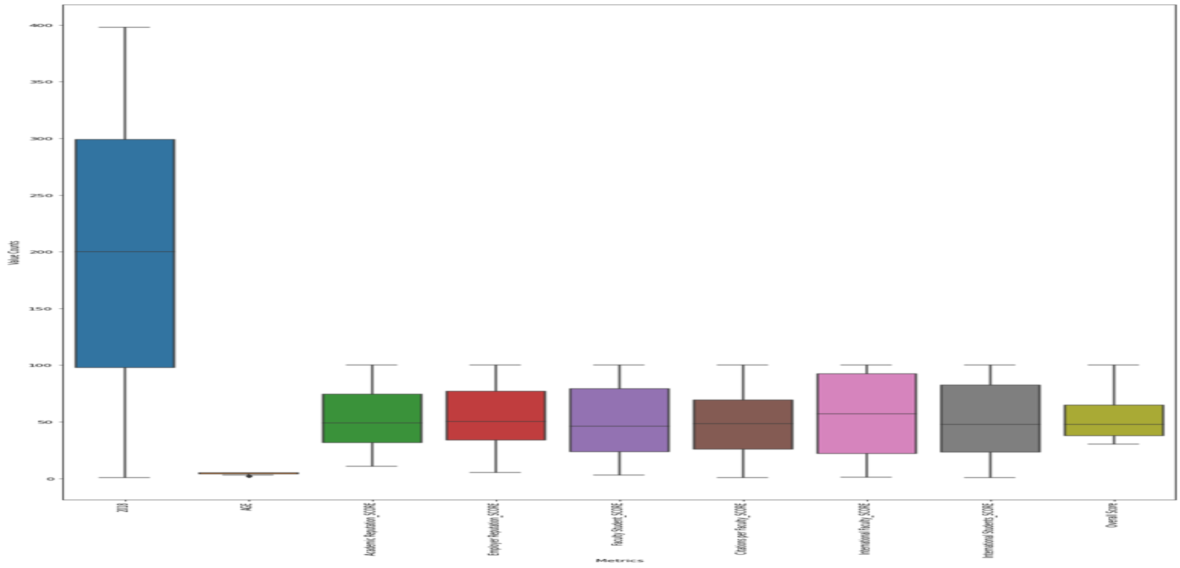


Figure 2: Box Plot Showing Outliers

In the 3-year dataset, the United States is the country with the highest number of universities. It is followed by the United Kingdom, then Germany, and least, Lithuania. See plot below:

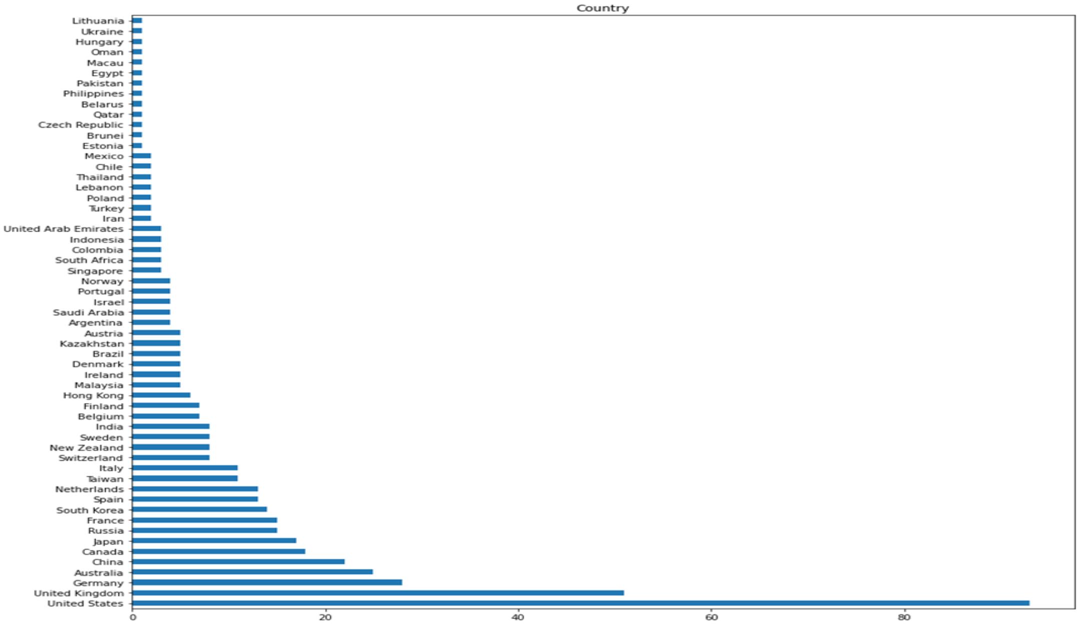


Figure 3: Universities Ranking by Country

Model Training, Evaluation, and Validation

a. **Model Training:** Since machine learning is all about learning the behaviour of the dataset and testing the behaviour with another dataset, the dataset was split into a training dataset and a testing dataset. The training dataset was used to build and validate the regression models.

b. **Model Evaluation:** The regression model evaluations were done using the regression metrics in the scikit-learn official documentation. Some of the metrics include: Mean Absolute Error (MAE), R-squares, Adjusted R-square, etc. The R-square is prominent among the metrics. However, the R-square score ranges from 0 to 1. When R-square is close to 1, the regression model is well-performing and vice versa. The R-square of negative indicates an erroneous model. The Mean Absolute Error was our main evaluation metric for the various regression models used for training the data.

c. According to our regression models in this study, the best-performing model was the CatBoost regressor.
d. **Model Validation:** There are many model validation techniques. The cross validation technique was used to validate our models.

Model Deployment

Hypertext Markup Language (Html) and Cascading Style Sheets (CSS) were used for the frontend. With a good model ready, Flask was used to deploy the model on Heroku in order to make live predictions. Here are the links;
Heroku: https://universityscoringengine.herokuapp.com/
GitHub: https://github.com/oadeniran/Hamoye-Team_deeplearning

Results

The important subject of forecasting to determine where to attend college has been addressed by this project. Expectedly, this will also motivate underachieving institutions to raise their performance standards and compete with top universities by taking into account the relevant variables shown in the analysis outcome.

Conclusion and Recommendation

The outcome of this investigation demonstrates that the quality of the dataset used to develop the model has a significant impact on the accuracy of the findings. Furthermore, the process of evaluating a university's overall score can be made possible by machine learning knowledge when used effectively. Therefore, the model should only be trained with accurate data from the previous year's records in order to get better outcomes.

Team Members

Oladimeji Williams	Emmanuel Ajibodu
Obiabuchi Nnanna	Emmanuel Bolarinwa
Jimoh Ridwan Adewale	Ogungbemi Oluwamayowa S.
Philip Odewole	Emmanuel Onuoha
Etieneabasi Kingsley Effiong	Kareem Jadesola
Isaac Tamunoboma Ayotamuno	Owolabi Adeniran
Ezerioha Ifeanyi Emmanuel	Imah Gift
Ajeyomi Adedoyin	Ugbo, Gregory Obinna
Muftaudeen Toyin Saheed	Paulina John
Udo Aniekan	