# Analyzing MS in Data Analytics and Data Science Programs: Developing A Prediction System for Program Ranking

KRANNERT SCHOOL OF MANAGEMENT
PURDUE UNIVERSITY

Jiacen Liu, Lingbo Fang, Songjun Jiang, Haodong Zhang, Sixing Luo, Matthew A. Lanham
Purdue University, Krannert School of Management
liu2480@purdue.edu; fang269@purdue.edu; jiang534@purdue.edu; zhan2874@purdue.edu; luo331@purdue.edu; lanhamm@purdue.edu

## ABSTRACT

This study analyzes current rankings and develops a new program ranking score for MS in Data Analytics and Data Science programs regarding institutional, educational, and environmental performance measures. Due to the variation in ranking methodologies used and often lack of transparency of data values submitted from those ranked institutions, we use predictive analytics to estimate what input levels are needed to achieve a certain rank for each ranking system. This study is novel in that programs can identify where their strengths and weaknesses lie with the goal of using their resources to make more strategic decisions to improve their future rankings.

## INTRODUCTION

During recent years, the global expansion of access to higher education has increased demand for information on academic quality. Therefore, ranking systems now play such a big role in shaping the opinions of current and potential students, parents, employers and government about the quality of tertiary education institutions. However, the emergence of this ranking obsession also causes a legitimate source of concern about its misuse, especially when it is used solely for promotional purposes (Marmolejo, 2015). Hence, without truly knowing what factors go into making these ranks, or how these factors were collected, and measured, it would be difficult to fairly judge these programs.

Figure 1. QS BA Masters Rankings 2021 Breakdown


**This study aims to answer the following research question:**

- Which ranking factor should have the highest indicator weight when calculating the overall program score?

## LITERATURE REVIEW

Unlike omnipresent studies conducted on national university rankings, studies focus on the ranking system for Data Analytics and Data Science programs are rare. Due to the lack of such studies for the programs, there hardly exists research that would identify the strengths and weaknesses by pinpointing the evaluation of program-specific indicators using RF and LR.

| Study | Summary |
|---|---|
| Jeremic et al. (2011) | Use the statistical I-Distance method to remedy the top 100 university ranking approach developed by ARWU. |
| Guarino et al.(2011) | Estimate the degree of uncertainty in the rankings and permits the assessment of statistically significant differences across universities. |
| Dobrota et al.(2015) | Use the composite I-distance indicator (CIDI) methodology to improve stability and reduced uncertainty of the QS ranking system. |
| Kapur et al. (2016) | Analyze how career outcomes of graduated students can affect the ranking of universities. |
| Roffo et al. (2017) | Propose a robust probabilistic latent graph-based feature selection algorithm that performs the ranking steps. |

## METHODOLOGY

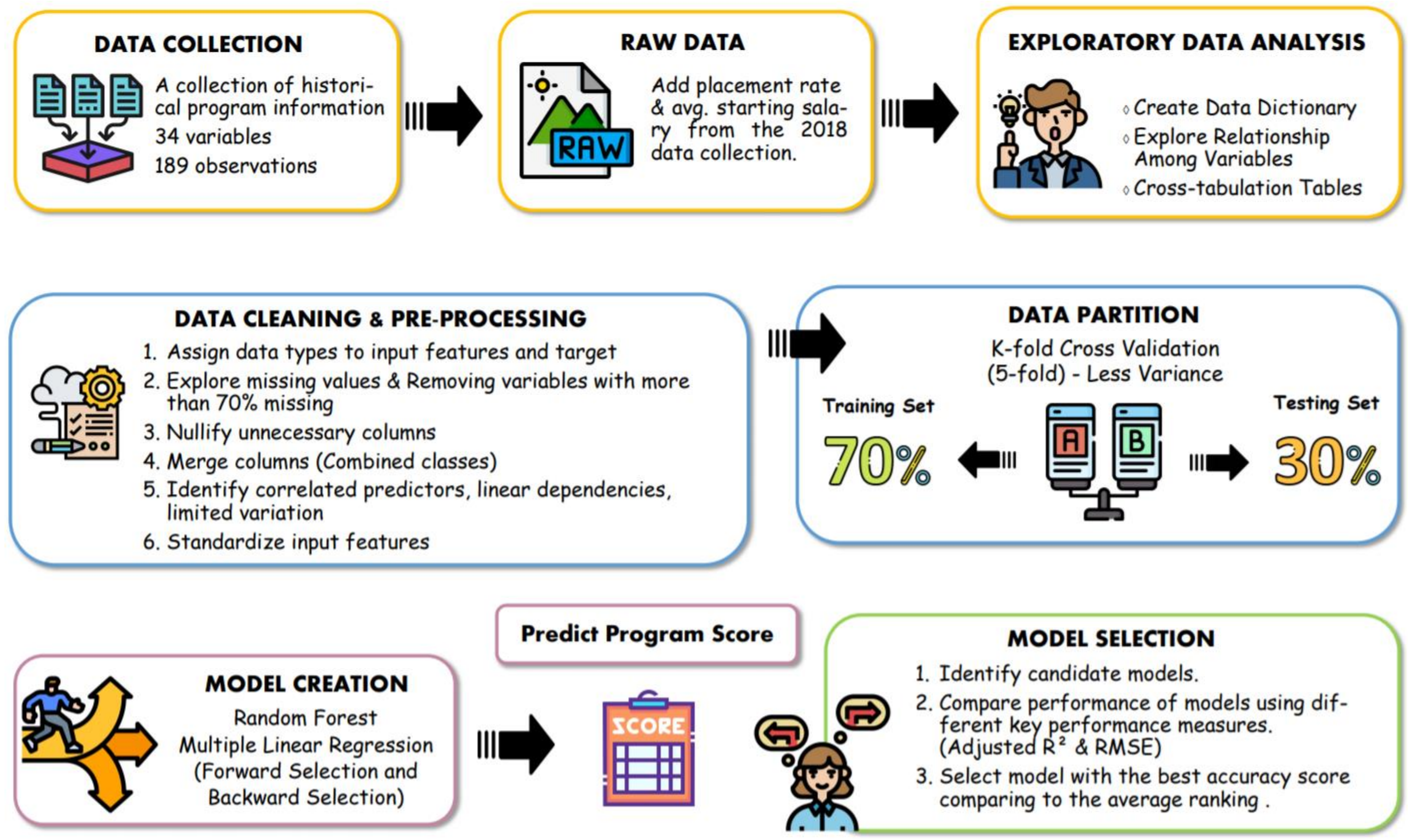
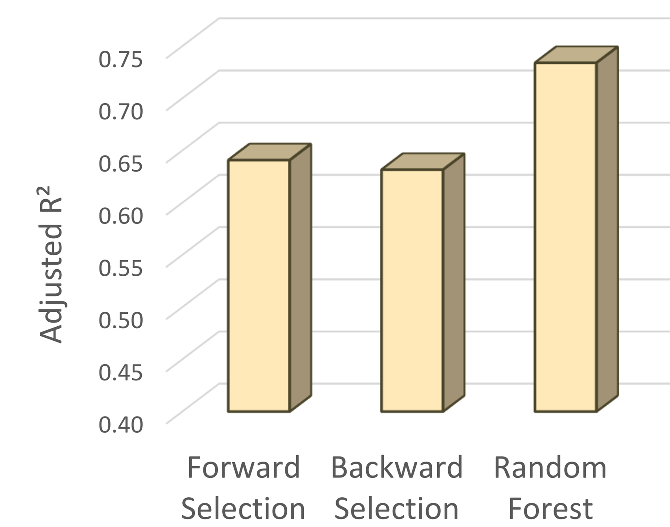



Figure 2. Study Design

### Model Selection

By running regressions through multiple variables, we finalized using random forest and linear regression model to determine good indicators in ranking MS in DS and DA programs. Random forest is preferred to reduce the variance, increase accuracy and less overfitting in decision trees. We used k-fold cv since it has less variance and train time than LOOCV. Models are evaluated with adjusted $R^2$ and comparing test set to train set to check model's generalizability.

## STATISTICAL RESULTS




Statistical Performance Measure: Adj. $R^2$

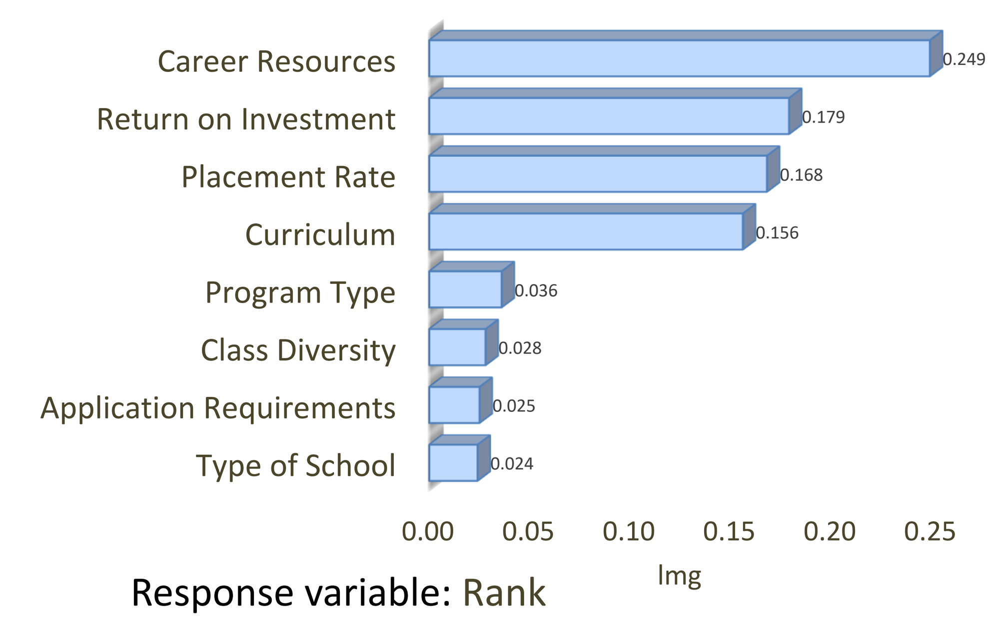
Candidate Model: Both RF and LR

*Difference between training and test set < 15%

Best Model: Random Forest

Since the adj $R^2$ is higher and lead to less error rate, we finally chose RF model.

Figure 3. Model Comparison Over Test Sets




Response variable: Rank

Proportion of variance explained by mode: 86.55%

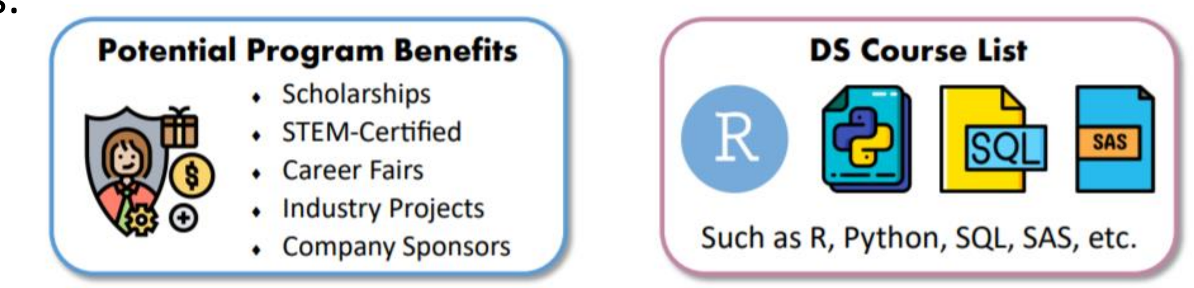
*Metrics are not normalized (rela = FALSE)

LMG calculates the relative contribution of each predictor to the R square with the consideration of the sequence of predictors appearing in the model.

Figure 4. Relative Importance of Predictors

## STUDY IMPLICATIONS

- **Career resources** has the highest relative importance between ranking indicators. Thus, it should have the highest indicator weight when calculating the overall program score.
- Program rankings can be sorted to show results for each of the performance indicators assessed.
- Both university per se and potential users such as companies, students, parents may utilize our ranking system to achieve their goals.




| University | Current Program Rank#: | Potential Improvement on Rank Indicator: | New Program Rank#: | Rank# Difference After Improvement: |
|---|---|---|---|---|
| | 11.64 | Number of Benefits: 3 +3 → 6 | 9.77 | ↑ 1.87 |
| BOSTON UNIVERSITY | 15.96 | Number of DS Courses: 3 +7 → 10 | 8.96 | ↑ 7.00 |

Figure 5. Potential Improvement on Rank Indicators

Universities may use data from rankings for analysis, strategic planning, and policy making. Companies can pick their qualified employees by expecting the ability that the student will equip out of the university after evaluating the ranking score of each master program.

## CONCLUSIONS

- Through our research, we have successfully developed a ranking system for MS Data Analytics and Data Science programs using R.
- Our ranking system is completely transparent which enables a quick detection of the weaknesses, strengths and opportunities for master programs.
- Weights of indicators were determined using the method developed by Lindemann, Merenda and Gold (LMG).
- **Career resources** should have the highest indicator weight when calculating the overall program score.
- In the future studies, more detailed information of each program may be added into the data set, including research publication impact, etc. Another future study on data collection can be obtaining required data from databases directly instead of limited web interface.

## ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham, Xinyu Wang, Theo Ginting, and our graduate student mentors Robyn Campbell and Judy Deng for their guidance and support on this project.