MONASH University

Identifying the Student Characteristics that Influence Employment Outcomes

Sia Chawla

Tran Dang Huy

Jiaqi Xu

MONASH
BUSINESS
SCHOOL

Department of Econometrics & Business Statistics

(03) 9903 4416
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

Report for
ETC5513 Group Assessment

2 June 2025

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Table of contents

1 Executive summary

This project examines which student experiences most strongly influence early career success, using a kaggle dataset from Shamim (2022). The analysis identified internships, high school GPA, field of study, networking skill, and university projects completed as the most significant predictors of job offers, while university ranking and soft skills demonstrated limited impact. The random forest model suggests that employers value a combination of practical experience, academic foundation, and selective engagement in co-curricular activities over institutional prestige. These findings challenge conventional assumptions about employability and provide meaningful insights to help students make informed, strategic career decisions.

2 Introduction (Jackie)

Graduates today face many questions about how their experiences influence their careers. Grades alone may not determine who gets the more job opportunities. This report explores which student experiences are associated with receiving more job offers.

To answer this question, we use a kaggle dataset from Shamim (2022) which contains 5,000 records from recent graduates. It includes information about students' academic background, personal demographic, and career-related outcomes. Rather than testing predefined theories, this project takes an open-ended, pattern-oriented approach. The aim is to explore which types of experiences appear most consistently linked to job outcomes and personal satisfaction.

3 Methodology

3.1 Data exploration

The data was collected from Kaggle and initially contains records for 5000 recent university graduates. During our data processing, we found 3.8% (192 rows) of the data have inconsistent gender values, as shown in Figure 1. Since this is likely to provide misleading information, we decided to exclude them from further analysis.

3.2 Model selection

Table 1 presents the first five records in the dataset and highlights the diversity of input variable types. The dataset includes numerical variables such as `SAT_Score`, categorical fields like `Field_of_Study`, and ordinal variables such as `Networking_Score`. This variety makes it unsuitable for models that require specific data type.

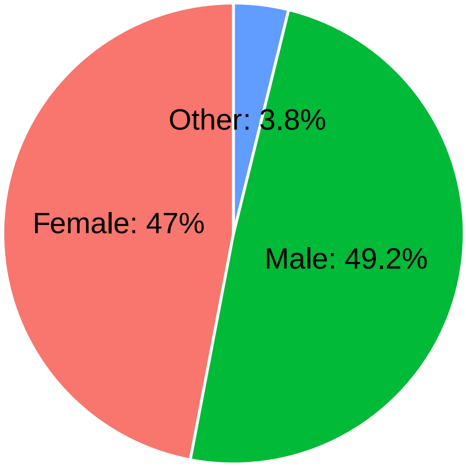

Figure 1: *Number of Gender Categories*

Table 1: *First 5 records in filtered dataset*

	1	2	3	4	5
Student_ID	S00001	S00003	S00004	S00005	S00006
Age	24	28	25	22	24
Gender	Male	Female	Male	Male	Male
High_School_GPA	3.58	3.42	2.43	2.08	2.40
SAT_Score	1052	1193	1497	1012	1600
University_Ranking	291	715	170	599	631
University_GPA	3.96	2.63	2.81	2.48	3.78
Field_of_Study	Arts	Medicine	Computer Science	Engineering	Law
Internships_Completed	3	4	3	4	2
Projects_Completed	7	8	9	6	3
Certifications	2	1	1	4	2
Soft_Skills_Score	9	1	10	10	2
Networking_Score	8	9	6	9	2
Job_Offers	5	0	1	4	1
Starting_Salary	27200	42400	57400	47600	68400
Career_Satisfaction	4	9	7	9	9
Years_to_Promotion	5	3	5	5	2
Current_Job_Level	Entry	Entry	Mid	Entry	Entry
Work_Life_Balance	7	7	5	2	8
Entrepreneurship	No	No	No	No	Yes

Figure 2 shows the distribution of the outcome variable, `Job_Offers`, which represents the number of job offers received. The distribution is relatively balanced across classes, meaning no class dominates the data, and no transformation is required. Furthermore, since `Job_Offers` is ordinal, it emphasises the use of classification model over regression model.

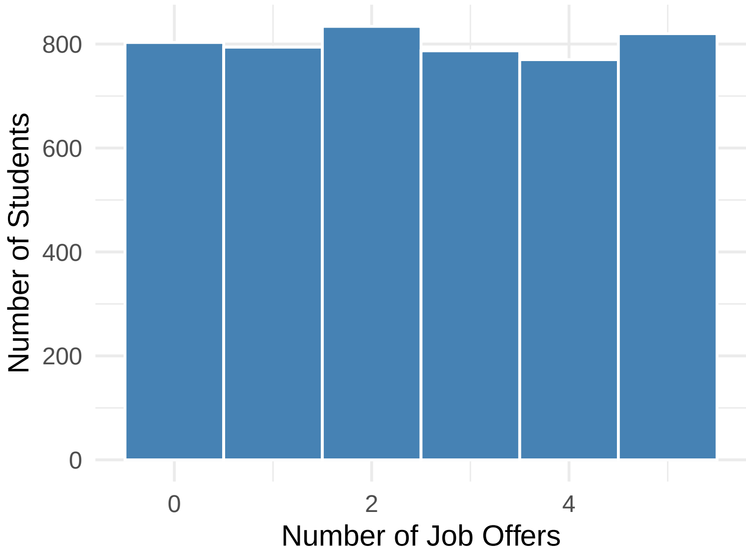

Figure 2: *Distributions of Outcome Variable*

Given these characteristics, random forest is an appropriate modelling choice. It is a flexible and robust algorithm that can naturally handle both numerical and categorical inputs without extensive preprocessing. In addition, it captures non-linear relationships and performs well even when input variables vary in scale or measurement type.

3.3 Accesing associations

After fitting the random forest model, we use permutation importance to rank all input variables (Figure 3), as referred by Gregorutti, Michel & Saint-Pierre (2017). This method measures how much the model's accuracy drops when the values of one variable are randomly shuffled. A larger drop means the variable was more important for the model's predictions. This allows us to assess which factors the model relies on most when estimating job offers.

The top 5 predictors were selected based on their importance scores (Figure 3) and further explored using boxplot with trend lines for the direction of association (**?@fig-trend**).

4 Results

4.1 Associated factors for receiving job offers.

Figure 3 ranks predictors by their importance in the random forest model. Students' practical experiences (Completed Internships and Academic Projects), academic background (High School GPA and Field of Study), and interpersonal networking skills are more associated with the number of job offers received. In contrast, factors such as University Ranking and Soft Skills Score appear to have minimal influence.

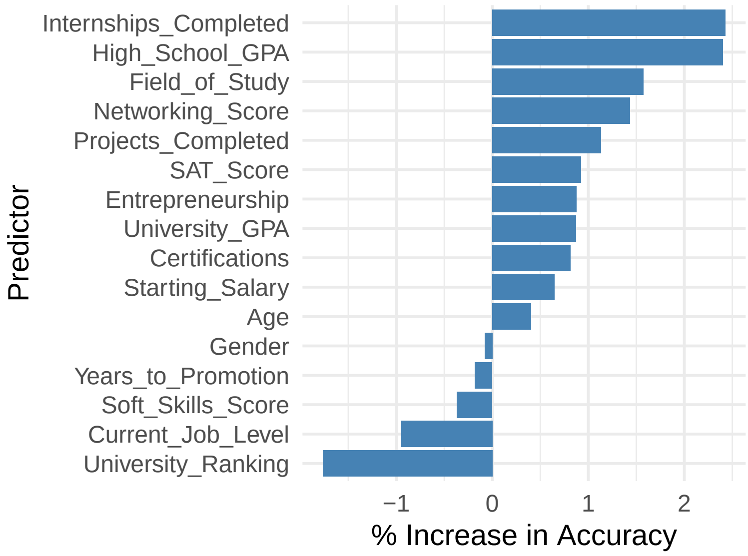

Figure 3: *Importance of Predictors for Number of Job Offers Received*

4.2 Association between important factors and job offers.

Figure 4 and Figure 5 further illustrate how job offers vary across the top five predictors.

The boxplot (Figure 4) shows that the median number of job offers falls between 2 and 3 for most categories. The plot also reveals a downward trend in medians as networking scores and project counts increase.

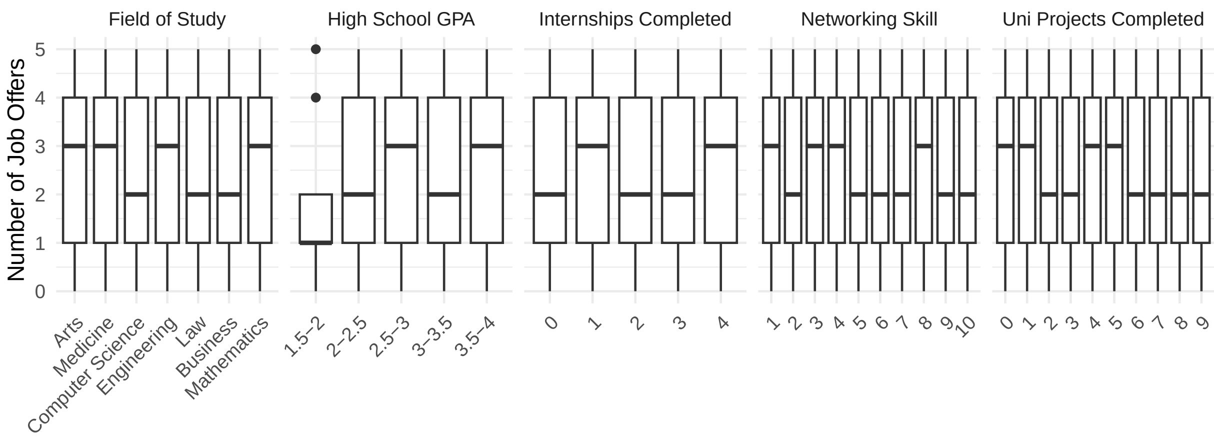

Figure 4: *Distribution of job offers across student categories*

While the patterns for high school GPA and internship experience are more subtle in the boxplot, Figure 5 shows clear upward trends, suggesting that students with higher GPAs and more internships receive more job offers on average.

Among fields of study, Medicine leads with the highest average job offers, followed by Arts and Mathematics, while Computer Science ranks the lowest.

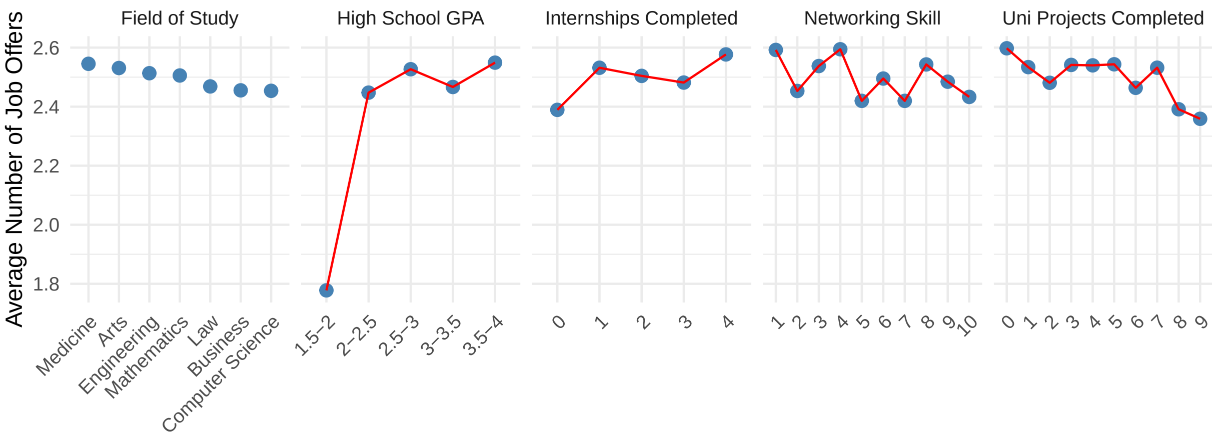

Figure 5: *Average number of job offers by predictor category*

5 Discussion, conclusion and recommendations

5.1 Discussion and Conclusion

The analysis revealed that internships completed, high school GPA, field of study, networking skill, and university projects completed were the most influential predictors of job offers. Internships and a strong academic foundation, particularly in high school, were consistently associated with better outcomes, indicating that employers value both preparedness and discipline. While field of study also shaped job offer patterns, the effect was relatively stable across disciplines. Interestingly, both networking skill and number of university projects showed a downward trend beyond moderate levels, suggesting that overinvestment in these areas may yield diminishing returns. These results highlight the nuanced ways in which various student experiences contribute to employability.

In conclusion, early career success is most strongly driven by a balance of practical experience, academic performance, and strategic engagement in co-curricular activities.

5.2 Recommendations

- **Students** are encouraged to pursue internships and maintain strong academic performance, particularly during high school, while engaging strategically in networking and project-based activities to avoid diminishing returns from overextension.
- **Higher education institutions** should embed structured experiential learning—such as industry placements and capstone projects—into academic programs, and ensure alignment between fields of study and evolving labour market demands.

- **Career development services** should provide targeted support that promotes a balanced portfolio of academic achievement, practical experience, and purposeful co-curricular involvement to optimise graduate employability.

Reference

Gregorutti, B, B Michel & P Saint-Pierre (2017). Correlation and variable importance in random forests. *Statistics and Computing* **27**(3), 659–678.

Liaw, A & M Wiener (2002). Classification and Regression by randomForest. *R News* **2**(3), 18–22.

Shamim, A (2022). *Education and Career Success*. https://www.kaggle.com/datasets/adilshamim8/education-and-career-success. Accessed: 2025-06-02.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, DD Vaughan, C Wilke, K Woo & H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Xie, Y (2025). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.50. `https://yihui.org/knitr/`.