Project 2 – Milestone 2

DSC 680

Alyssa Graham

Business Problem

Graduate employability is a top priority for most colleges and universities, yet they often lack clarity on which student attributes help students find job placements. This project aims to build a predictive model to identify the most important factors that affect a students likelihood of being places in a job after graduation.

Background/ History

Over the last decade, student job placement rates have become a key performance indicator for colleges and universities. Students face growing competition while employers increasingly seek both technical knowledge and soft skills. This has prompted many schools to consider data-driven approaches so they can do their best to support student success and prepare them for their careers.

Data Explanation

The dataset used for this analysis was taken from Kaggle and includes simulated records from 10,000 different students. Some notable variables represented in this dataset include; College_ID, IQ, Prev_Sem_Result (GPA from the previous semester), CGPA (cumulative GPA), Academic_performance, Intership_Expereince, Extra_Curricular_Score, Communication_Skills, Projects_Completed, and Placement. Before the analysis, all features were checked for missing values, and categorical variables were encoded to enable their use with machine learning models. This final cleaned dataset was used for the analysis.

Methods

Before starting the analysis, the dataset was cleaned using step previously stated. An Exploratory Data Analysis was then conducted to better understand feature relationships using tools like correlation matrices and distribution plots. The dataset was then split into training and testing sets using an 80/20 ratio. A Random Forest Classifier was selected for its ability to handle both numerical and categorical values. Model performance was then evaluated using a confusion matrix, accuracy, score, precision, recall, and F1-Score.

Analysis

The predictive model was trained using a Random Forest model which achieved approximately 87% accuracy on the test set. Key predictors for this placement include; Cumulative GPA, Internship Experience, and Projects Completed. Students who had internship experience and had a higher overall GPA were seen to be significantly more likely to be placed into a job out of college. While not quite a impactful, communication skills and extracurricular scores also contributed to the likelihood of job placement.

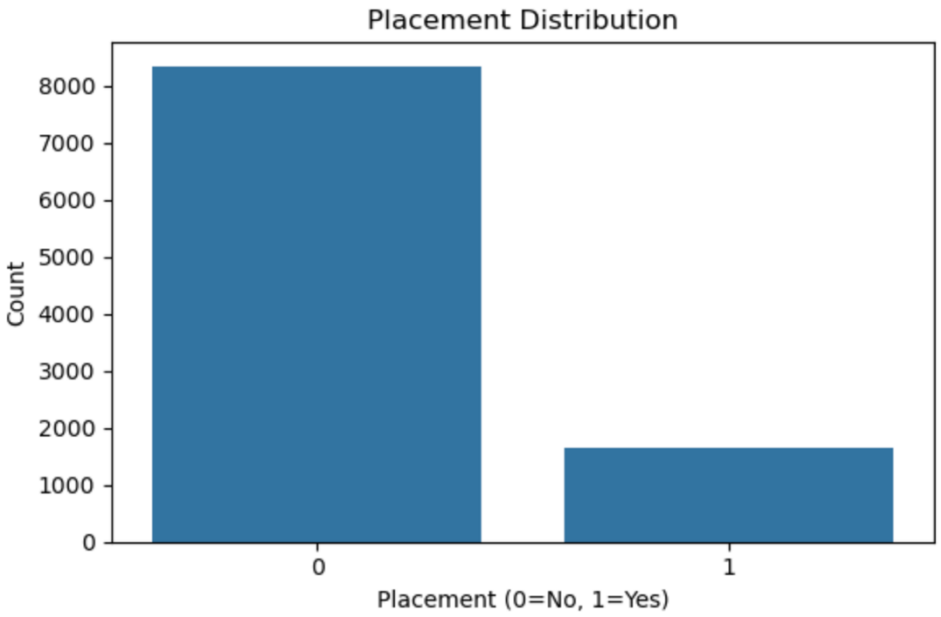

Figure 1: Placement Distribution Plot

Figure 1 shows the basic distribution of students who received a placement (1) and students who did not receive a placement (0). We see from this that there are significantly less students who had jobs after graduating compared to those who didn't. This makes those jobs highly competitive and students must make themselves stand out in order to get one of these positions.

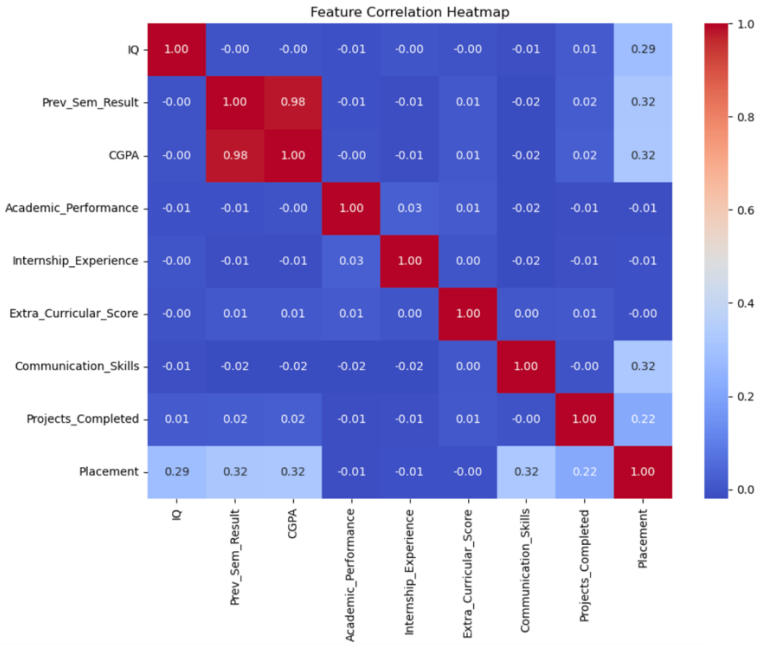

Figure 2: Feature Correlation Heatmap

A correlation heatmap was generated to understand the relationships between features prior to modeling. In Figure 2, we see a strong positive correlation between Cumulative GPA and Previous Semester GPA. This suggests that there was consistency in academic performance among students. We also see that Internship Experience showed a noticeable correlation with Placement, our target variable. This reinforced the assumption that real world experience contributes to job acquisition.

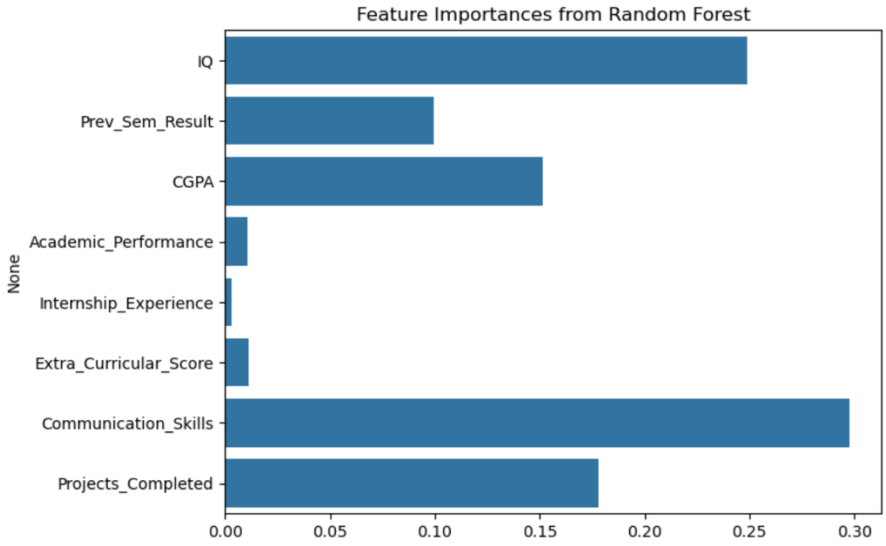

Figure 3: Feature Importance Chart

Figure 3 shows the feature importance chart derived from the Random Forest model. We see that Cumulative GPA, Internship Experience, and Projects completed rank as the top three predictors of job placement. These features showed significantly higher importance scores compared to other variables such as IQ or Academic Performance. This suggests that hands-on experience and sustained academic success is more valued by employers that test scores alone.

```
--- Accuracy Score ---
1.0

--- Classification Report ---
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1674
           1       1.00      1.00      1.00       326

    accuracy                           1.00      2000
   macro avg       1.00      1.00      1.00      2000
weighted avg       1.00      1.00      1.00      2000


--- Confusion Matrix ---
[[1674    0]
 [   0  326]]
```

Figure 4: Confusion Matrix

A confusion matrix was used to evaluate prediction accuracy across both classes (students who were placed and students who were not placed), we see this in Figure 4. The model demonstrates a balanced performance that minimizes both false positives (predicting placement where there was none) and false negatives (missing students who should be flagged for support). This confusion matrix provides evidence that the classifier is accurate, and it is practical for use in real-world decision making.

Together, these analyses and visuals validate the model's performance and highlight impactful traits that institutions can support to improve placement outcome.

Conclusion

This predictive model confirms that academic and experiential variables, specifically internship involvement and cumulative performance, are strong indicators of students being placed in a job. Institutions can used these models to flag students who might benefit from career preparation and educational counseling services. This analysis supports the idea that early interventions, such as offering internships and coaching communication, can improve placement outcomes.

Assumptions

This analysis assumes that the input data is accurate and the recorded placement status reflects actual job acquisition. Additionally, it presumes that the provided features are consistently measured across colleges and universities.

Limitations

There are some limitations to consider with this analysis. The dataset does not include industry specific or role specific placement outcomes. This may limit how precisely one can interpret placement quality, results may vary by programs/ roles. Additionally, demographic variables such as gender or socioeconomic background are not included, which restricts the model's ability to assess fairness or potential bias. Finally, because this is a simulated data set, results may not generalize direct real-world college settings. Results conducted with real-world data may show varying results.

Challenges

One major challenge with this analysis was addressing the imbalance in placement labels since significantly more students were not placed in roles. Additionally, interpreting encoding categorical data in meaningful ways for stakeholders was important for transparency. Another complication was discovered when ensuring the model remained interpretable while maintaining high performance.

Future Uses/ Additional Applications

The most obvious use is for colleges to help students by providing coaching with variables that help with job placement. Beyond this, this framework can be adapted to track academic progress, predict dropout risk, or personalize learning interventions. Future versions of this analysis could include time-series data across semesters or even NLP analysis of student resumes. Integrating these inputs from learning management systems could enhance prediction and responsiveness.

Recommendations

Institutions should invest in internship programs and focus on developing communication and project-based learning skills. A predictive model like the one I developed here should not be used as a standalone tool rather, it should be used as part of a larger ecosystem that includes academic advising and career services. Regular audits and validations should be conducted to ensure fairness and to adapt the model to keep up with evolving trends.

Implementation Plan

To implement these ideas, institutions should first generate their own model using their own data. The model can then be integrated to academic platforms for use by academic advisors. The staff should be trained on how to interpret model predictions and how to intervene ethically. Over time, the model will need to be monitored and retrained annually to incorporate the most recent outcomes and keep information up to date.

Ethical Assessment

Ethical concerns include potential bias if the model favors students from specific colleges or academic backgrounds. It is also important to ensure transparency in how predictions are used. Students must be able to understand that the model is a support tool. Privacy should be a top priority, with student data being made anonymous and kept protected at all stages. Lastly, the model should be assessed regularly to check for fairness.

References

National Center for Education Statistics. (2023). *Fast facts: Employment outcomes of college graduates*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/fastfacts/display.asp?id=40

Sumedh. (2022). *College student placement* [Data set]. Kaggle. https://www.kaggle.com/datasets/sumedh1507/college-student-placement

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Audience Questions:

1. How realistic is it to apply a simulated dataset to real-world placement strategies?
2. What is the main advantage of using machine learning over traditional statistical methods for this problem?
3. Why were certain features like gender or socioeconomic status not included in the dataset?
4. Was there a feature that you expected to be influential that wasn't?
5. Which feature turned out to be the most predictive of placement? Why?
6. How can you address potential class imbalance in placement outcomes?
7. What steps were taken to validate the model's accuracy?
8. How would you ensure transparency and fairness if this model were deployed institutionally?
9. What are some specific actions colleges could take based on the model's results?
10. What are some ways to collect additional real-world data to improve predictions?