THE COMPLEXITY OF SUCCESS

Prepared for
Phuong (Kem) Nguyen-Le, Ph.D.

Prepared by
Samuel S. Muma
University of Maryland

May 6, 2025

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

A student excels by every academic measure — top grades, competitive internship, and leadership roles. On paper, they are the blueprint of success, the student universities showcase and the candidate every company should want. Yet, come graduation, they're overlooked. Meanwhile, another student, with average grades but hands-on experience, certifications, and a few well-placed connections, lands their dream job and accelerates quickly.

These contrasting paths raise a critical question: Can academic performance, extracurriculars, and internships really predict career success? As a junior data scientist, I set out to answer that — not with opinion, but with data. Sourcing the *Education & Career Success* dataset from Kaggle, it recorded 5,000 students' academic records, work experience, and early career outcomes.

For a holistic view of career success, I engineered a Composite Career Success Score combining salary, job offers, satisfaction, and rate of promotion. I then trained models, linear regression, gradient boosting, and neural networks, to detect patterns and make predictions. At first, the results across all three models were exciting (MAE ≈ 0.15, MSE ≈ 0.04, $R^2 \approx 0.45$), showcasing minimal errors in predictions, hinting at a well fitted model for real-world decision making. However, further examination revealed data leakage – a variable used in both the input and the target – which artificially inflated performance.

Once corrected, R^2 dropped significantly, revealing that the features available did not strongly explain variance in career outcomes. Yet the models continued to maintain low prediction error rates, indicating they could still make reasonable average predictions despite weak underlying relationships.

These findings led to an insightful realization: good model metrics don't always mean we understand the system. Real-world success is often shaped by factors the data cannot capture. Timing, opportunity, personality, and luck. This report follows the path from early optimism to critical insight, emphasizing both the potential and the limits of machine learning in modeling human trajectories.

2. PROBLEM

Career success is shaped by a myriad of factors — Job offers, starting salary, career satisfaction, work life balance, entrepreneurial endeavours and even the less tangible elements such as timing and personal connections. Yet, students and job seekers often face uncertainty about which of these factors will have the greatest impact on their future. Academic institutions, career advisors with our career center, and employers similarly seek to understand what truly drives successful career outcomes, aiming to guide emerging talent more effectively.

In a world increasingly reliant on data-driven decision making, can machine learning uncover patterns that explain and predict career success? If academic achievements, internships, and networking efforts could reliably forecast outcomes like job offers, starting salaries, and career satisfaction, it would empower individuals to make strategic choices and allow institutions to tailor support more effectively.

This project investigates these questions by applying predictive analysis to an *'Education & Career Success'* dataset generated synthetically using data from real world education and career trends. Encompassing university rankings, GPAs, soft skills, and early career outcomes. Through the development of advanced machine learning models, the project seeks to determine whether measurable educational and extracurricular experiences can predict holistic career success — not only in terms of salary, but also job opportunities, satisfaction, and promotion speed.

The business and social value of this work lies in its potential to revolutionize how we view academic and career preparation. If successful, such models could help students focus their efforts on the most important aspects of their academic career. *The University Career Center & The President's Promise* can further provide evidence-based guidance. Employers, who partner with Universities, can refine talent identification beyond surface-level metrics.

However, as this study will reveal, the complexity of human outcomes challenges the boundaries of algorithmic prediction, raising important questions about the role of data in understanding personal and professional growth.

3. METHODOLOGY

3.1 Data Preparation

The initial stage of the project involved a structured exploration of the sourced dataset, which contained 5,000 student records including academic, extracurricular, and career outcome information. A data dictionary was created during this stage to define and understand each feature's role within the modeling process. This early documentation served to ensure clarity around data usage, identify potential target variables, and confirm any assumptions before any transformations were conducted

Table 1. Data Dictionary for Education & Career Success Dataset

Feature Name	Description	Data Type	Possible Values
Student_ID	Unique identifier for each student	String	S00001
Age	Age of the student	Integer	18-30
Gender	Gender identity of a student	String	Male, Female, or Other
High_School_GPA	High school GPA	Float	2.0 - 4.0
SAT_Score	Standardized test score	Integer	900 - 1600
University_Ranking	Ranking of the university attended	Integer	1-1000
University_GPA	University GPA	Float	2.0 - 4.0
Field_of_Study	Major or discipline	String	Computer Science
Internships_Completed	Number of internships completed	Integer	0 - 4

Projects_Completed	Number of personal/academic projects completed	Integer	0 - 9
Certifications	Number of additional certifications earned	Integer	0 - 5
Soft_Skills_Score	Soft skills rating	Integer	1 - 10
Networking_Score	Score based on professional networking and connections	Integer	1 - 10
Job_Offers	Number of job offers received after graduation	Integer	0 - 5
Starting_Salary	First job salary in USD	Integer	150,000
Career_Satisfaction	Career satisfaction level	Integer	0 - 10
Years_to_Promotion	Time taken to receive the first promotion	Integer	1 - 5
Current_Job_Level	Career level	String	Senior
Work_Life Balance	Work-life balance rating	Integer	1 - 10
Entrepreneurship	Whether the individual started a business	Boolean	Yes/No

The dataset was imported using pandas.read_csv(). Basic information regarding data types, non-null counts, and overall structure was reviewed using df.info(). This verified the integrity and structure of the incoming data. An initial preview of the data was performed using df.head(), providing insight into feature layout and example records. The complete set of features and their descriptions is summarized in Figure 1.

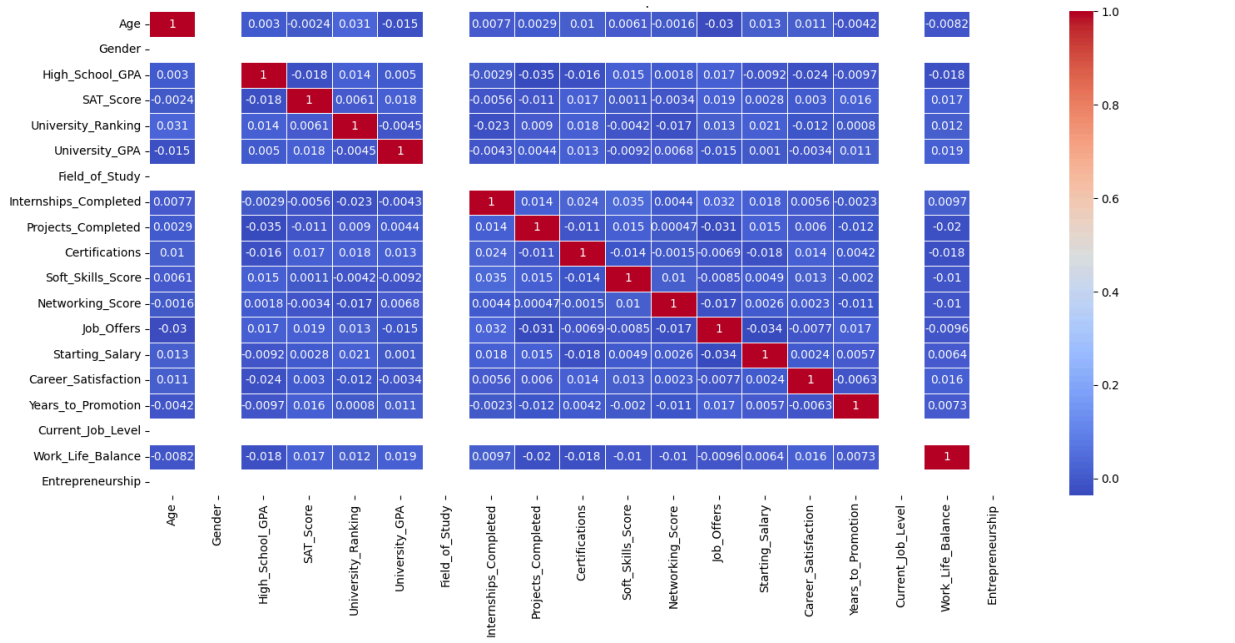
3.2 Feature Engineering

Following the initial data preparation, feature engineering was undertaken to refine the dataset and create meaningful variables to support the modeling phase. The Student_ID column was removed as it did not have any predictive value. A check for missing values was performed with df.isnull().sum(), revealing minimal missing values across the dataset.

Missing numerical values were imputed with the median of each respective column, ensuring robustness against outliers. Categorical variables were filled with the placeholder “Unknown” to maintain data integrity, reducing biases. They were then transformed using one-hot encoding, with the first category dropped to prevent multicollinearity. This conversion enabled the machine learning algorithms to process categorical information numerically.

A correlation heatmap (Figure 1) was generated to assess the strength of linear relationships among the numerical features. The results revealed that most academic, extracurricular, and early career variables were only weakly correlated, with few values exceeding even modest thresholds. This early insight suggested that individual features alone would not strongly explain variance in career success. As a result, new interaction terms were engineered to combine conceptually related indicators (e.g., GPA × internship count), in an effort to surface latent patterns that might not appear in isolated metrics.

Figure 1. Correlation heatmap of numerical variables



All numerical features were scaled using MinMaxScaler to normalize the range between 0 and 1. This step was crucial for models sensitive to feature magnitude differences. Variance Inflation Factor (VIF) scores were computed for numerical features to assess any hidden multicollinearity within the dataset. The VIF values confirmed there was no multicollinearity present. Several new composite scores were engineered to better capture multidimensional aspects of a student's performance:

- **Academic Performance Score:** Combined University_Ranking (40%) and Univeristy_GPA (60%) to represent academic achievement.
- **Extracurricular Score:** Weighted combination of Internships_Completed (50%), Projects_Completed (50%), Soft_Skills_Score (30%), and Networking_Score (20%) to reflect overall extracurricular engagement and social capital.
- **Career Success Score:** Integrated Starting_Salary (50%), Job_Offers (50%), Career_Satisfaction (30%), and an inverted measure of Years_to_Promotion (20%) to reflect early career success.[1]

[1] After engineering composite features, the original contributing variables were dropped to prevent redundancy. The cleaned dataset was saved as a CSV file for subsequent modeling, ensuring a clean, structured input for machine learning workflows.

4. RESULTS

4.1 Model Selection

To evaluate the extent to which academic performance, extracurricular activities, and other measurable factors predict career success, three machine learning models were developed and tested: Lasso Regression, Gradient Boosting Regressor, and a Neural Network built using PyTorch. Each model was trained to predict a scaled composite variable; Career_Success_Score_Scaled, which was generated via MinMaxScaler transformation to normalize the target for regression analysis. The final set of features used in modeling included a combination of engineered composite scores and encoded categorical indicators:

- Academic_Performance
- Extracurricular_Score
- Work_Life_Balance
- Field_of_Study (one-hot encoded for Business, Computer Science, Engineering, Law, Mathematics, and Medicine)
- Entrepreneur (binary)

The data was split into training and testing sets (80/20 ratio), and all features were standardized using StandardScaler to ensure equal weighting across inputs during model training.

4.2 Lasso Regression

A Lasso Regression model was trained using 5-fold cross-validation (LassoCV) to identify the optimal regularization parameter (alpha). Lasso was selected for its ability to perform feature selection by penalizing less relevant variables, reducing overfitting. After training, the model's predictions were evaluated on the test set using standard regression metrics. The trained model was exported to save for reproducibility.

4.3 Gradient Boosting Regressor

A Gradient Boosting Regressor was trained using GridSearchCV to optimize hyperparameters including n_estimators, learning_rate, and max depth. This tree-based ensemble method was chosen for its ability to model non-linear relationships and reduce bias through additive training. After identifying the best-performing parameter set through cross-validation, the model was evaluated on the test set and saved for future inference.

4.4 PyTorch Neural Network

To explore a more dynamic, non-linear modeling architecture, a feedforward neural network was implemented using PyTorch. The model architecture consisted of:

- Input layer with units matching the number of features
- Two hidden layers with ReLU activation and dropout (to prevent overfitting)
- Final output layer with a single neuron for regression

The network was trained using Mean Squared loss and the Adam Optimizer, with early stopping based on validation loss to prevent overfitting. A training loop was executed for up to 100 epochs, with the best-performing model weights saved using checkpointing. Input features and targets were converted to tensors, and separate DataLoaders were constructed for training and validation using mini-batches. Upon training completion, the model was evaluated on the test using MAE, MSE, and R2 metrics.[2]

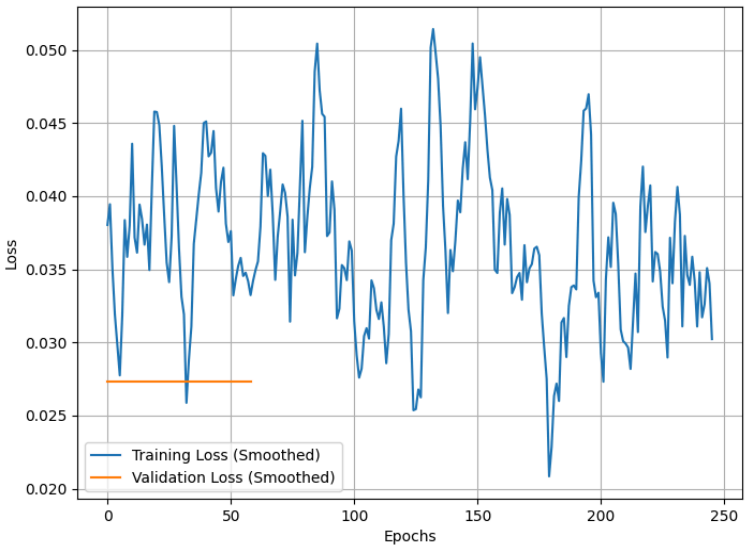
Table 2. Predictive Accuracy Across Modeling Approaches

Model	MAE	MSE	R^2
Lasso Regression	≈ 0.15	≈ 0.04	≈ 0.00
Gradient Boosting	≈ 0.15	≈ 0.04	≈ -0.01
Neural Network	≈ 0.15	≈ 0.04	≈ -0.02

Despite low prediction error, the R^2 score for all models hovered around 0.00, indicating that the models did not meaningfully explain the variance in the target variable. In other words, while the models' predictions were, on average, within a certain range of the actual values, they struggled to uncover the underlying relationships that drive career outcomes. A limitation rooted not in the models themselves, but the data's structure.

The learning curve for the best-performing neural network model, achieved after extensive hyperparameter tuning and feature refinement, is presented in Figure 2.

Figure 2. Learning Curve of the Neural Network Model



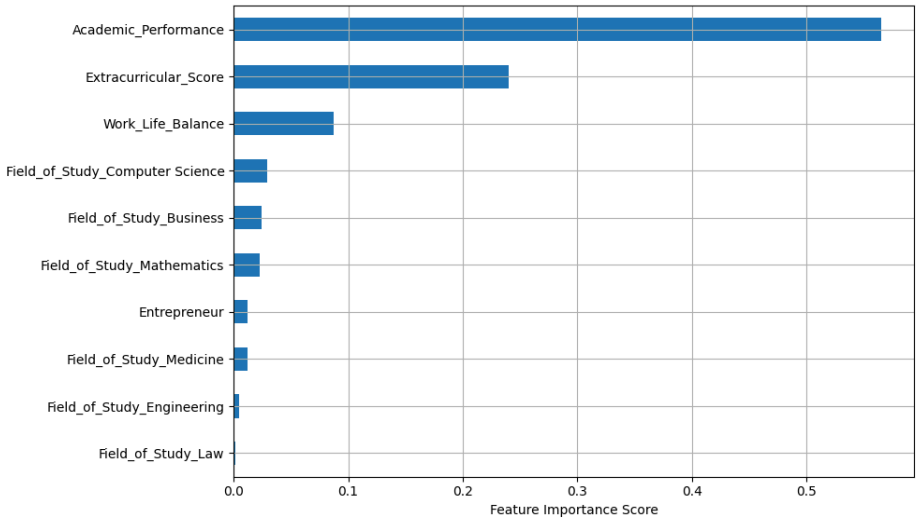
The training loss shows an initial decrease, but then exhibits significant fluctuations throughout the training process, failing to converge to a consistently low value. More critically, the validation loss remains relatively flat and consistently higher than the training loss after the initial epochs. The challenges in achieving a better learning curve likely stem from the inherently weak linear relationship observed in the numerical features (averaging a correlation of 0.02 with the target variables).

[2] All models, Lasso Regression, Gradient Boosting, and the PyTorch Neural Network, were saved to disk in serialized format (.joblib for scikit-learn models and .pth for PyTorch).

While the creation of composite features for Academic Performance, Extracurricular Engagement, and indicators of early career interest (e.g., Field_of_Study, Work_Life_Balance) is aimed to capture more nuanced aspects and potentially noise, the resulting learning curve still indicates poor generalization. This suggests that even these more holistic measures of academic background might not be strong predictors of the chosen career outcome variable, Career Success Score, which is influenced by a multitude of factors beyond academic performance, could also contribute to the difficulty in establishing strong predictive relationships.

To gain insights into the relative influence of different factors on career outcomes, I examined the feature importance scores derived from the Gradient Boosting Regressor model. FIgure 3 displays these values, with the length of each bar representing the relative contribution of each feature to the model's predictive ability.

Figure 3. Feature Importance from Gradient Boosting Regressor



The Academic_Performance composite score emerged as the most important feature by a significant margin. This suggests that the model heavily relied on the integrated measure of university ranking and GPA when making predictions regarding career outcomes.[3]

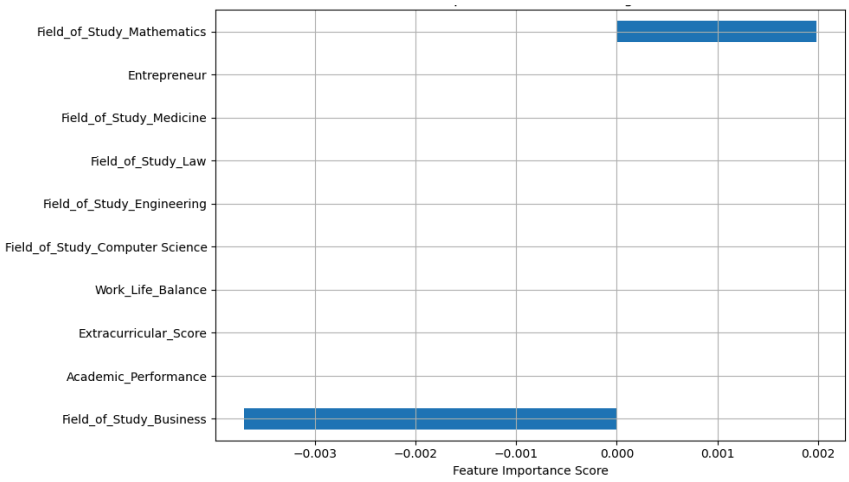
The Extracurricular_Score, a composite of internships, projects, soft skills, and networking activities, showed a moderate feature importance (≈ 0.23) within the model. While not the dominant driver, its influence remains notable. This result reflects on a broader reality in early career outcomes: while employers often prioritize GPA and institutional prestige as screening criteria, hands-on experience and interpersonal effectiveness play a crucial role in differentiating candidates beyond their resume.

Interestingly, the Work_Life_Balance feature demonstrated a small but non-negligible influence on predicted career outcomes, indicating that perceived balance may play a subtle role in early success trajectories. In contrast, individual Field_of_Study indicators all recorded importance scores below 0.1, suggesting that once academic performance and extracurricular engagement were accounted for, the specific major contributed minimally to the model's predictions.

[3] Within the composite metric, university ranking (60%) carries more influence than GPA (40%) in shaping career outcomes, highlighting the potential impact of institutional prestige in early career trajectories.

Similarly, the binary Entrepreneur feature had a minimal impact on model predictions, indicating that entrepreneurial status alone did not significantly influence early career outcomes when considered alongside academic and extracurricular factors. I also examined feature importance using a Lasso Regression model, which performs automatic feature selection by shrinking the coefficients of less than important features to zero. The resulting scores are visualized in Figure 4.

Figure 4. Feature Importance from Lasso Regression



The Lasso Regression model identified only two features with non-zero coefficients. Field_of_Study_Business and Field_of_Study_Mathematics. The negative coefficient for Business (≈ -0.003) suggests a very weak inverse linear relationship with the target variable, while the positive coefficient for Mathematics (≈ 0.002) indicates a slightly favourable linear relationship.

Crucially, the Lasso model effectively reduced the coefficients of all other features, including the composite scores for Academic_Performance and Extracurricular_Score, as well as Work_Life_Balance and Entrepreneur to zero. This implies that within the constraints of a linear model with an L1 penalty, these features were not deemed to have a significant independent linear relationship with the target variable compared to the selected fields of study.

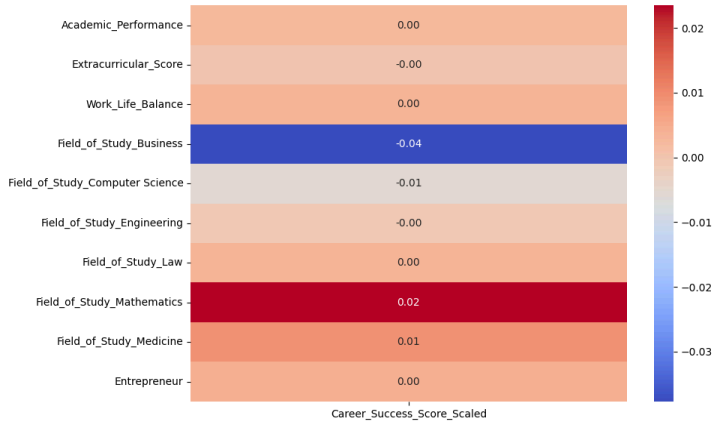
The stark contrast in feature importance between the Lasso Regression and Gradient Boosting models underscores how model choice influences what patterns are prioritized. Lasso, with its built-in preference for sparsity, seeks the simplest possible linear explanation. In this case, it was determined that only the fields of Business and Mathematics contributed marginally to predicting career outcomes. This reflects Lasso's underlying assumption: that relationships are linear and additive.

In contrast, Gradient Boosting does not penalize feature inclusion in the same way and can capture non-linear, multi-feature interactions, which likely explains its broader distribution of feature importance. The Lasso model's narrow focus highlights the limitations of linear models in complex, multifactorial human data, where simplicity comes at the expense of nuance.

5. APPLICATIONS

The insights from this project offer tangible value to NGOs and educational institutions supporting students in their transition from education to employment. While the predictive models were unable to fully explain individual career trajectories ($R^2 \approx 0.00$), they consistently achieved low error rates (MAE ≈ 0.15, MSE ≈ 0.04), meaning they can still be used to flag general trends and risk indicators across large student cohorts. For example, analysis of Figure 5 (reveals that traditional academic and extracurricular metrics show very weak linear relationships with early career outcomes, while certain fields of study exhibit slightly more influence.

Figure 5. Inputs vs. Composite Career Success Score



Notably, Business majors show a weak negative correlation with the composite success score, whereas Mathematics majors show a weak positive correlation. For an NGO or university career center, this suggests that a broad, GPA-centered approach may overlook critical nuances. Instead, interventions could be more targeted by field, offering specialized advising, career guidance, or professional development support where students may face comparatively lower outcomes.

Quantitatively, if such models were used to prioritize outreach or programming for even 10% of a 5,000-student population, organizations could reallocate resources more efficiently and potentially prevent career stagnation or underemployment in hundreds of cases annually. This approach does not replace human guidance but rather augments it, enabling case managers, advisors, or program directors to scale their impact using evidence-based flags and scoring systems.That said, deploying these insights at scale introduces both technical and organizational challenges.

From a technical perspective, the lack of strong feature-target relationships limits model precision — emphasizing the need for richer, more granular data, such as student goals, mentorship access, or socioeconomic background. Organizationally, success will depend on cross-functional buy-in, clear communication, and a commitment to ethical implementation. Predictive tools must be used not to limit opportunity, but to enhance equity by identifying and supporting underserved or overlooked student groups. With responsible integration, predictive modeling can be a powerful tool in advancing the mission of equitable, data-informed career support.

6. LIMITATIONS

While this project offers meaningful insights into patterns of early career success, it also highlights several ethical considerations and limitations that must be acknowledged.

First, the dataset itself reflects only the variables that were synthetically recorded, such as GPA, internships, and field of study, while omitting critical contextual factors like socioeconomic status, race, disability status, access to mentorship, or systemic barriers. This raises the possibility of algorithmic bias: if the underlying data reflects historical inequalities or blind spots, the models may unintentionally reinforce those gaps, especially when used in decision-making contexts (e.g., advising, funding, or intervention allocation).

Additionally, while the models performed consistently in terms of MAE and MSE, they exhibited very low R^2 scores, indicating that most variance in career success remains unexplained. This suggests that the model's predictions are only useful at the aggregate level, and should never be used to make definite judgements about individual potential. In edge cases — such as students with atypical pathways, unmeasured strengths, or nontraditional definitions of success — the models are likely to underrepresented their true value.

From a fairness standpoint, certain features like field of study may correlate with outcomes for reasons not visible in the data — such as systemic hiring biases in certain industries. Without adjusting for those external factors, there's a risk that models may reflect correlation without understanding the cause, which can lead to misleading conclusions if not interpreted with care. To mitigate these risks, predictive tools like the ones developed here should only be deployed with strong human oversight, transparent communication about model limitations, and ongoing evaluation.

Furthermore, institutions should work toward collecting more inclusive, intersectional data, and regularly audit model outputs across demographic groups to ensure equitable performance. Ultimately, this project is not intended to define success, but to support its discovery — and that requires ethical humility, a commitment to equity, and recognition that the most important predictors may be the ones we haven't yet measured.

WORKS CITED

University of Maryland. (n.d.). *Job trends & student outcomes*. University of Maryland. Retrieved February 23, 2025, from https://careers.umd.edu/explore-careers/industries-career-paths/job-trends-student-outcomes

Georgetown University Center on Education and the Workforce. (n.d.). *Home*. Georgetown University. Retrieved February 23, 2025, from https://cew.georgetown.edu/

Muma, S. (2025). Predicting Career Success [Source Code]. GitHub. https://github.com/iamsamuelm/predicting_career_success

PeopleScout. (n.d.). *Predictive analytics: The talent tool for success*. PeopleScout. Retrieved February 23, 2025, from https://www.peoplescout.com/insights/predictive-analytics-talent-tool/#:~:text=For%20instance%2C%20when%20filling%20certain,the%20success%20of%20a%20candidate.

Shamim, A. (2023). *Education & Career Success* [Data set]. Kaggle. Retrieved February 23, 2025, from https://www.kaggle.com/datasets/adilshamim8/education-and-career-success

Frost, J. (2018). *How to interpret r-squared in regression analysis*. Statistics by Jim. https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Wilson, M., Robertson, P., Cruickshank, P., & Gkatzia, D. (2022). *Opportunities and risks in the use of AI in career development practice*. Journal of the National Institute for Career Education and Counselling, 48(1), 48–57. https://doi.org/10.20856/jnicec.4807

Musoro, J. Z., Zwinderman, A. H., Puhan, M. A., ter Riet, G., & Geskus, R. B. (2014). *Validation of prediction models based on lasso regression with multiply imputed data.* BMC Medical Research Methodology, 14(1). https://doi.org/10.1186/1471-2288-14-116

Natekin, A., & Knoll, A. (2013). *Gradient boosting machines, a tutorial*. Frontiers in Neurorobotics, 7(21). https://doi.org/10.3389/fnbot.2013.00021

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., Devito, Z., Raison Nabla, M., Tejani, A., Chilamkurthy, S., Ai, Q., Steiner, B., & Facebook, L. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Caulfield, J. (2024, January 17). *APA Format for Tables and Figures | Annotated Examples*. Scribbr. Retrieved May 5, 2025, from https://www.scribbr.com/apa-style/tables-and-figures/