Exploring Key Influences on Post-Ph.D. Employment Choices: A Data-Driven Binary Classification Approach

Aayushi Verma
M.S. Data Science, Pace University
Advisor: Dr. Christelle Scharff

Abstract

After obtaining a Ph.D. degree, there are many career paths open to graduates. Existing literature analyze these post-Ph.D. career trends and turnover using statistical methods and social science frameworks, but there remains a gap in leveraging data science methods to address these questions. In this study, we employ data science and machine learning models to analyze the IPUMS Higher Ed survey dataset spanning 2003 - 2013, focusing on the U.S. STEM workforce. Our analysis centers on understanding the factors influencing Ph.D. graduates in choosing between academia and industry. We utilize six binary classification models to predict the employment sector of these graduates and employ feature engineering methods to identify the characteristics influencing their decision. This research contributes to the field by demonstrating applications of computational methods, particularly machine learning, to augment the research methodology.

Keywords: machine learning; binary classification; feature extraction; feature engineering; exploratory data analysis; data science

Research Question

Can we predict the employment sector (academia vs. industry) of Ph.D. degree holders and identify the key characteristics influencing this career choice?

Literature Review

Most studies in the field use frameworks and statistical models to explain job satisfaction (Dorenkamp & Weiß, 2017), intention to leave the higher education sector (Dorenkamp & Weiß, 2017, Szromek & Wolniak, 2020), turnover rates (White-Lewis et al., 2022, Xu, 2008), etc.
Very few studies use machine learning techniques to predict attributes that cause turnover (Birzniece et al., 2022).
Another study (Makridis, 2021) that used the IPUMS Higher Ed dataset also used statistical models to perform their analysis.
Sheetal et al., 2022 uses a similar longitudinal survey dataset and apply machine learning models like XGBoost to find the most important predictor variables.
Dorenkamp & Weiß, 2017 found that postdocs tend to leave academia because of effort-reward imbalance and professional overcommitment, Szromek & Wolniak, 2020 found that researchers are dissatisfied because of lack of recognition, increased workload without compensation, and more. White-Lewis et al., 2022 found that researchers leave because of faculty and tenure status factors. Xu, 2008 found that women are more likely to leave STEM academic careers due to reasons like dissatisfaction with research support, advancement opportunities, and free expression of ideas.

Dataset

IPUMS Higher Education is a longitudinal survey dataset covering the U.S. STEM workforce from 2003 to 2013 in discrete intervals, comprising 531,216 rows and 126 columns. We wrote a script to create a data dictionary for the coded feature names and values. We filtered the dataset to only include respondents who have earned Ph.D. degrees, who were full-time employed, and who work in either academia (defined as a 4-year college/medical institution), or industry. We performed an extensive investigation into the features and their correlations to downsize the dataset size, given its multidimensionality.

The choice of employment sector for PhD graduates strongly depends on **teaching activity**, **employer size**, **federal support**, and **salary**.

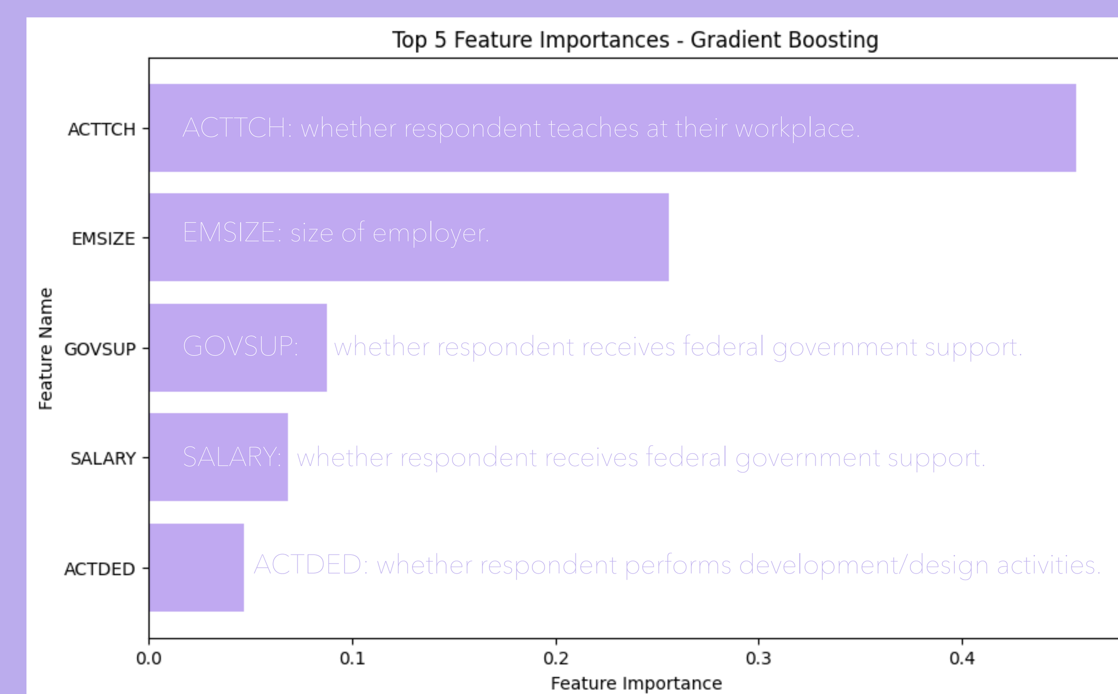


Fig. 1. Key features ranked by importance, determined through the Gradient Boosting method.

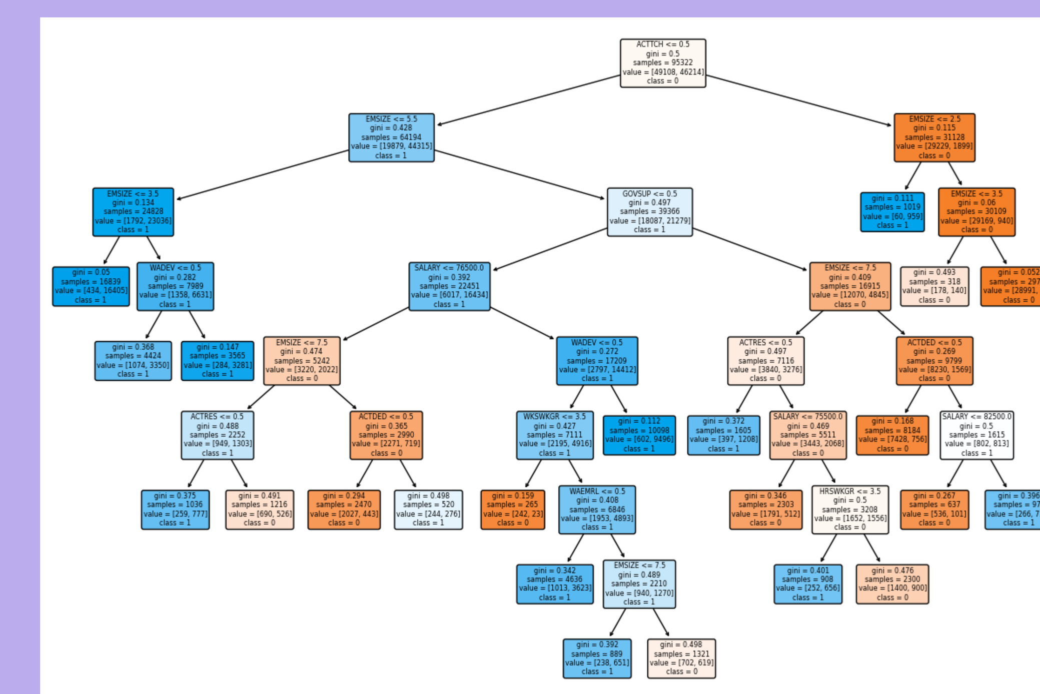
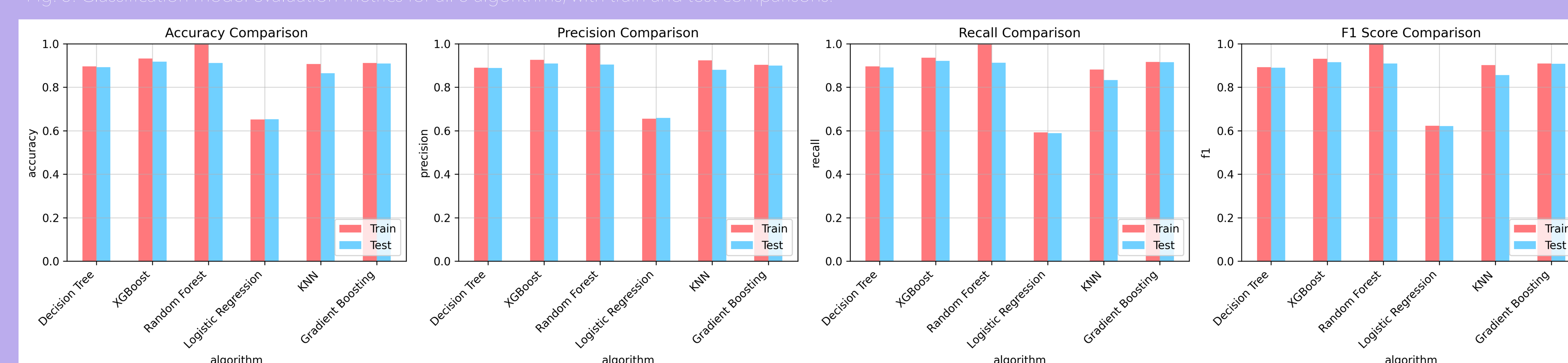


Fig. 2. Decision Tree visualized.

Fig. 3. Classification model evaluation metrics for all 6 algorithms, with train and test comparisons.



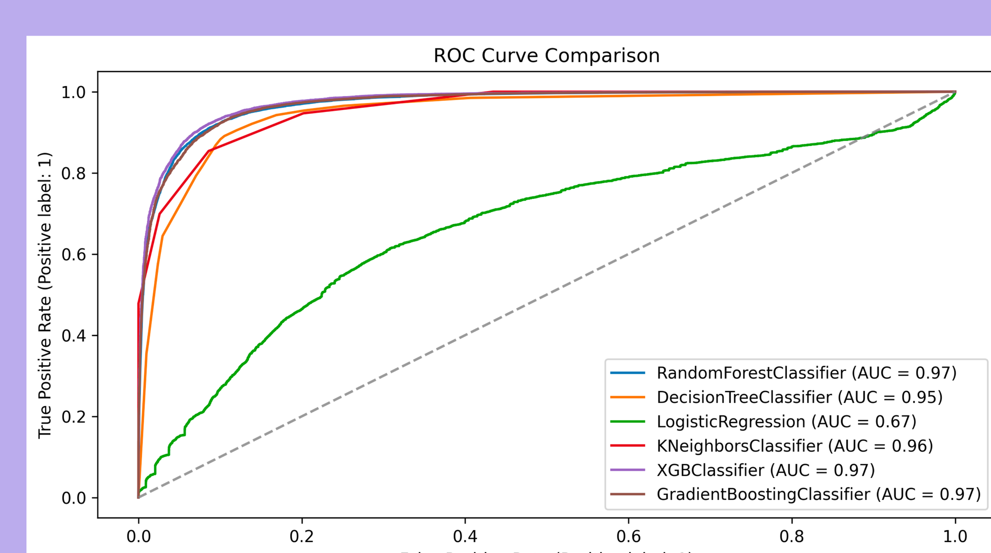


Fig. 4. Receiver Operator Characteristic (ROC) curve and Area Under Curve (AUC) scores for all 6 algorithms. Logistic Regression is the worst performer.

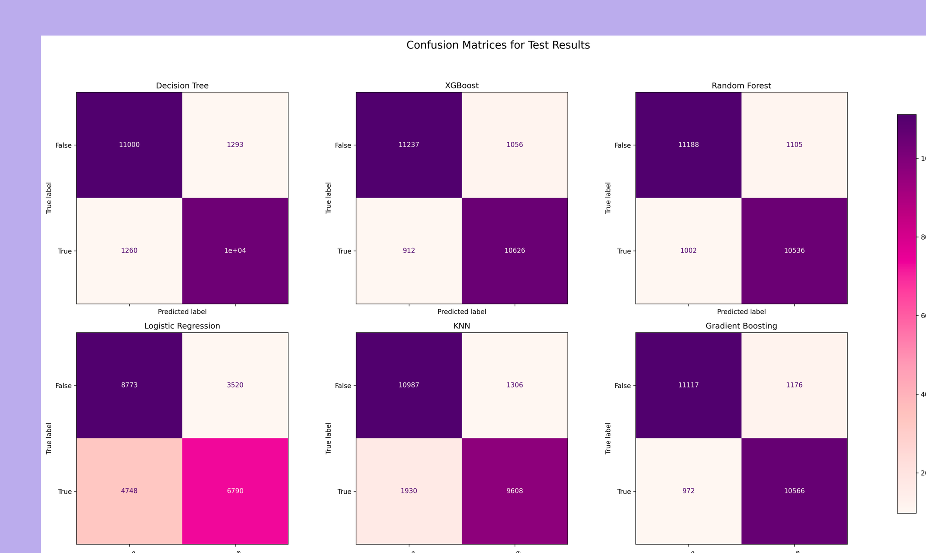


Fig. 5. Confusion matrices for all 6 algorithms. Logistic Regression is the worst performer.

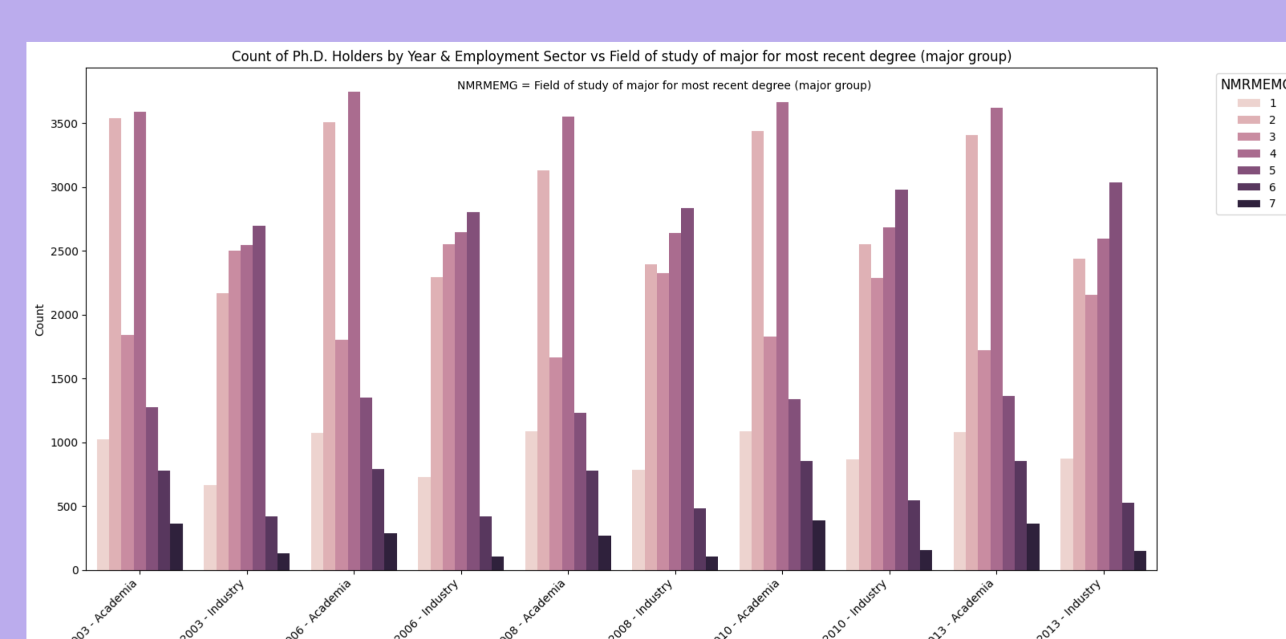


Fig. 6. An example of the EDA process. Each feature, in this case NMRMEMG (field of major), was visualized by year and employment sector, to investigate any trends, and to determine whether to discard the feature. Downsampling was necessary due to the high dimensionality of the dataset.

The code repository for this research project.

Methodology

- Performed extensive exploratory data analysis (EDA) to understand the data - checked for null and missing values, correlations of features, and ensured data was numerical.
- Identified target variable (EMSEC) and relevant features for answering research question and filtered data accordingly (MRDG = 3, EMSEC = (2,4), LFSTAT = 1, WRKG = 1).
- Examined each of 126 variables by year and employer sector, used statistical correlations and domain knowledge, discarded 103 features for a final dataset size of (119153 x 23).
- Split dataset by 80% train and 20% test.
- Identified 7 binary classification algorithms to classify our dataset using our target variable (Decision Tree, XGBoost, Support Vector Machine, Random Forest, Logistic Regression, k-Nearest Neighbors, Gradient Boosting).
- Obtained feature importances for each algorithm, evaluated best-performing models using metrics (confusion matrices, accuracy, precision, recall, F1 score, ROC-AUC score).

Results

Due to hardware constraints, we could not run the Support Vector Machine model. Regardless, we ran the other 6 models to predict the class of EMSEC as either 'industry' or 'academia'. Based on extensive analysis of our performance metrics defined earlier, and using domain knowledge as guidance, we chose two primary evaluation metrics: accuracy and ROC/AUC score. We found the top three models, in order, to be: Gradient Boosting, XGBoost, and Decision Tree. From all three models' feature importances, we found the top three factors in classification as either industry or academia is teaching activity, employer size, and federal government support. Other factors include salary and development activity. Due to hardware constraints, we were not able to perform any hyperparameter tuning.

Algorithm	Train Accuracy	Test Accuracy	AUC Score
kNN	0.907	0.864	0.975
Gradient Boosting	0.912	0.910	0.971
Decision Tree	0.896	0.893	0.970
XGBoost	0.933	0.917	0.948
Logistic Regression	0.652	0.653	0.926
Random Forest	0.997	0.912	0.665

Conclusion & Future Work

We used the longitudinal IPUMS Higher Ed survey dataset to analyze employment trends of PhD graduates, and factors that influence their decisions to work in either industry or academia. We performed a comprehensive investigation of the data and its features to reduce its dimensionality, and used six binary classification algorithms to predict survey respondents' employment sectors as either industry or academia. We extracted the most relevant features pertaining to each class, and determined that teaching activity, employer size, and federal support were the most important divisors of either class. We demonstrated the usage of data science and machine learning models to solve a problem in the social science domain, as compared to traditional statistical analyses. Limitations include implicit feature boundaries rather than explicit decision boundaries, lack of hyperparameter tuning due to hardware constraints, and lack of using statistical frameworks in literature to supplement analysis. Additionally the findings are based on the dataset features – there are many real-world factors that are not accounted for due to the survey design. Future work includes robustly finding the most optimal model with hyperparameter tuning, fine-tuning the feature engineering and extraction process, and using these data science methods in conjunction with the existing statistical methods and social science frameworks.

References

I. Dorenkamp and E.-E. Weiß, "What makes them leave? A path model of postdocs' intentions to leave academia," High Educ, vol. 75, no. 5, pp. 747–767, May 2018, doi: 10.1007/s10734-017-0164-7.
A. R. Szromek and R. Wolniak, "Job Satisfaction and Problems among Academic Staff in Higher Education," Sustainability, vol. 12, no. 12, Art. no. 12, Jan. 2020, doi: 10.3390/su12124865.
D. K. White-Lewis, K. O'Meara, K. Mathews, and N. Havey, "Leaving the Institution or Leaving the Academy? Analyzing the Factors that Faculty Weigh in Actual Departure Decisions," Res High Educ, vol. 64, no. 3, pp. 473–494, May 2023, doi: 10.1007/s11162-022-09712-9.
Y. J. Xu, "Gender Disparity in STEM Disciplines: A Study of Faculty Attrition and Turnover Intentions," Res High Educ, vol. 49, no. 7, pp. 607–624, Nov. 2008, doi: 10.1007/s11162-008-9097-4.
C. A. Makridis, "(Why) Is There a Public/Private Pay Gap?," Journal of Government and Economics, vol. 1, p. 100002, Mar. 2021, doi: 10.1016/j.jge.2021.100002.
I. Birzniece, I. Andersone, A. Nikitenko, and L. Zvirbule, "Predictive Modeling of HR Dynamics Using Machine Learning," in 2022 7th International Conference on Machine Learning Technologies (ICMLT), Rome Italy: ACM, Mar. 2022, pp. 17–23. doi: 10.1145/3529399.3529403.
R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," Journal of Applied Science and Technology Trends, vol. 1, no. 2, Art. no. 2, May 2020, doi: 10.38094/jastt1224.
A. Sheetal, Z. Jiang, and L. Di Milia, "Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers," Applied Psychology, vol. 72, no. 3, pp. 1339–1364, 2023, doi: 10.1111/apps.12435.