Predicting Minority STEM Attrition: Leveraging Machine Learning for Retention Policy-making

Md Ohiul Islam
University of Nevada, Reno

Research Objective

This study applies machine learning methods to a substantial dataset (N=488,160) to find predictive models and factors influencing STEM field exit patterns among minority workers. Black and Hispanic STEM degree-holders especially females exit STEM at a higher rate. It aims to model minority college graduates' decisions to leave their specialized fields. The analysis reveals that policymakers can leverage big data and machine learning techniques, including shrinkage methods, ridge regression, LASSO, regression trees, etc. to precisely predict STEM exit patterns. This approach allows for the identification of consistent factors influencing STEM exit among minority workers, assesses the heterogeneity of exit patterns across different minority groups, and aids in the design of targeted policies to reduce the STEM exit rate. The findings suggest that a comprehensive analysis using a variety of machine learning models can offer valuable insights for policy formulation aimed at retaining minority workers in STEM fields.

Methods and Key Variables

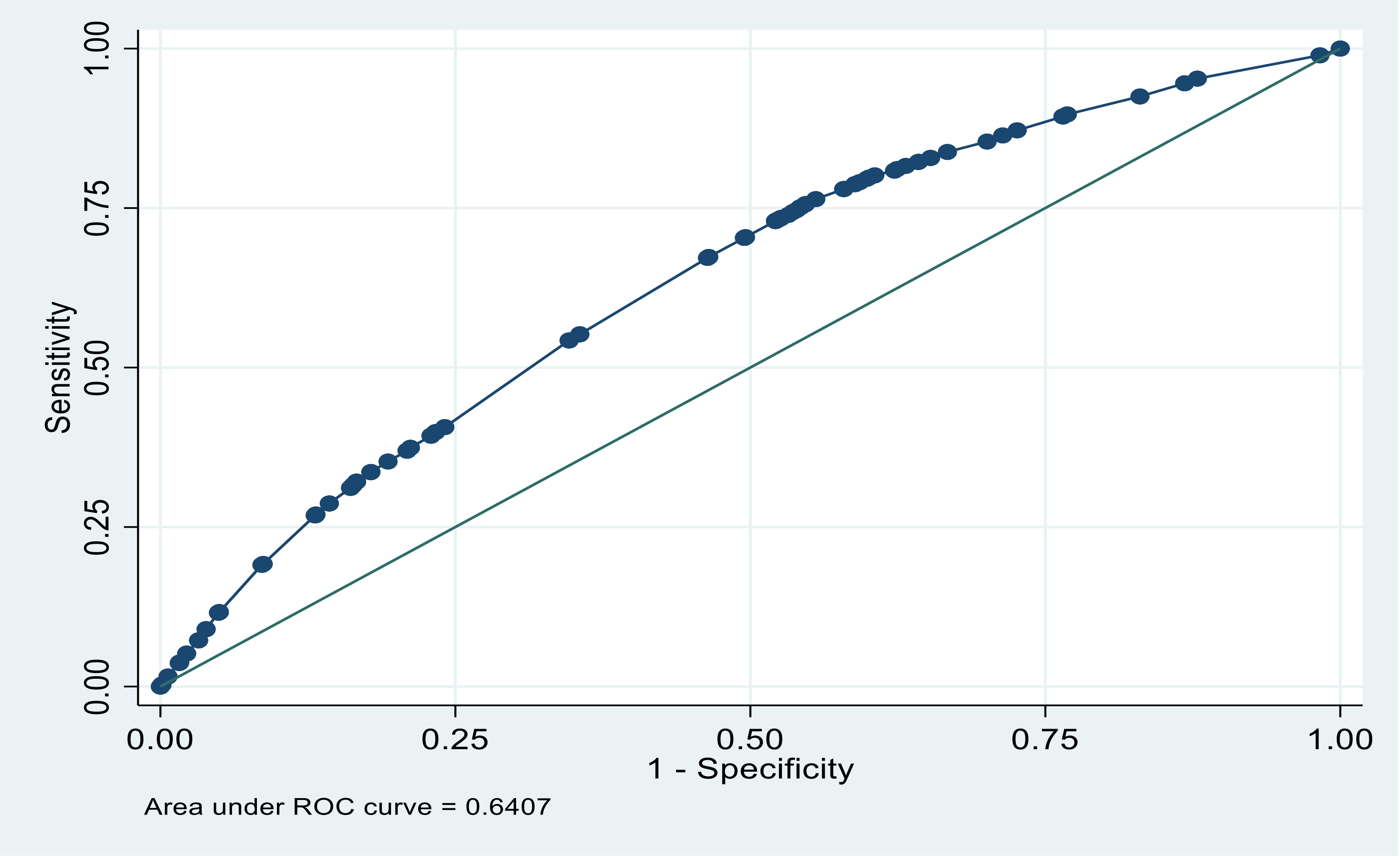
1.OLS (Ordinary Least Squares): A linear regression method that estimates model parameters by minimizing the sum of squared differences between observed and predicted values.
2.Probit: A regression model that estimates the probability of a binary outcome using the cumulative standard normal distribution to model the dependent variable.
3.Logit (Logistic Regression): A regression analysis that models the probability of a binary dependent variable using the logistic function.
4.LDA (Linear Discriminant Analysis): A classification method that projects features onto a lower-dimensional space to maximize class separability based on linear combinations of predictors.
5.QDA (Quadratic Discriminant Analysis): Similar to LDA but allows for non-linear separation between classes by using quadratic decision boundaries.
6.Ridge Regression: A linear regression technique that includes an L2 penalty on the size of coefficients to address multicollinearity and prevent overfitting.
7.Lasso (Least Absolute Shrinkage and Selection Operator): A regression analysis method that applies an L1 penalty to reduce the number of variables in a model by shrinking some coefficients to zero.
8.Elastic Net: A linear regression model that combines L1 and L2 regularization penalties to benefit from both Lasso's variable selection and Ridge's multicollinearity handling.
9.Regression Trees: A non-linear model that divides the predictor space into distinct regions, making predictions based on the mean outcome within each region for regression tasks.

Variables

Response variable: 1[STEM College Graduate working in a non-STEM job]
Main Explanatory Variables of Interest: Primary and Secondary activities Indicators, Parental Educational Achievement, Marital Status, Number of Children, Degree Field, Fields of Specialization, and Demographic Indicators

Results

Figure 1. Logistic ROC Curve



10-fold Cross-validated Logistic Regression generates Area under the (ROC) curve between 0.64-0.65

Table 1. Model Error Rates

	Training Error Rate	Testing Error Rate	Testing Error Rate	Testing Error Rate	Testing Error Rate
OLS	0.12	0.12	0.12	0.12	0.12
LOGIT	0.12	0.12	0.12	0.12	0.12
PROBIT	0.12	0.12	0.12	0.12	0.12
LDA	0.13	0.13	0.13	0.13	0.13
QDA	0.51	0.50	0.50	0.50	0.50
Samples	**Full**	**Full**	**Black**	**Hispanic**	**Female**
Cross-validation	✗	✓	✓	✓	✓

Table 2. LASSO: 10-fold cross-validation with 100 lambdas

Description	Lambda	Out-of-sample Dev. Ratio	CV Mean Deviance
First λ	.0037173	0.0337	.6958259
λ Before Selected λ	. 0000469	0.0354	.6945844
	.0000427	0.0354	.6945843

Figure 2. Cross-validation in Ridge Regression to Choose Optimal Lambda

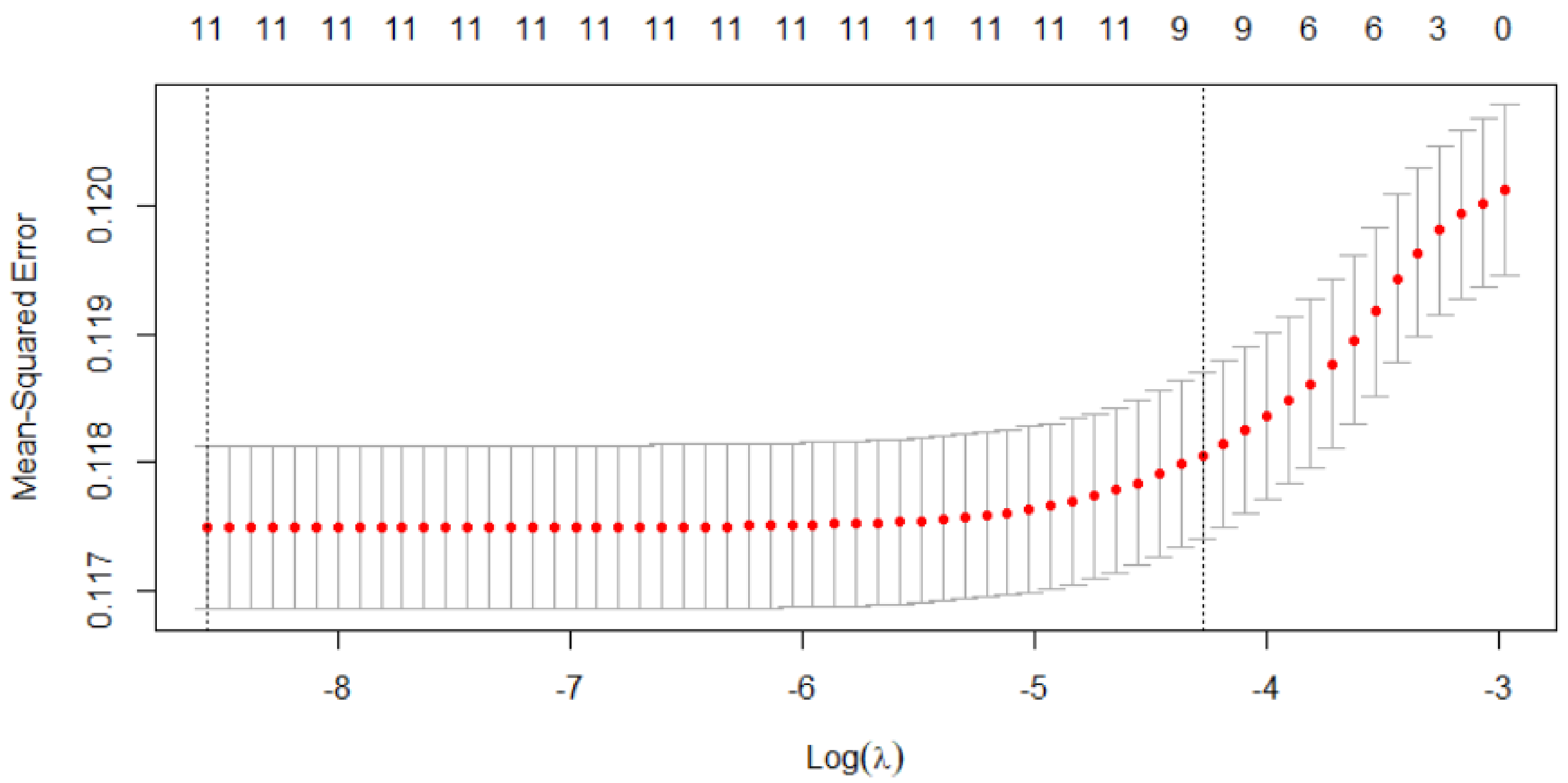
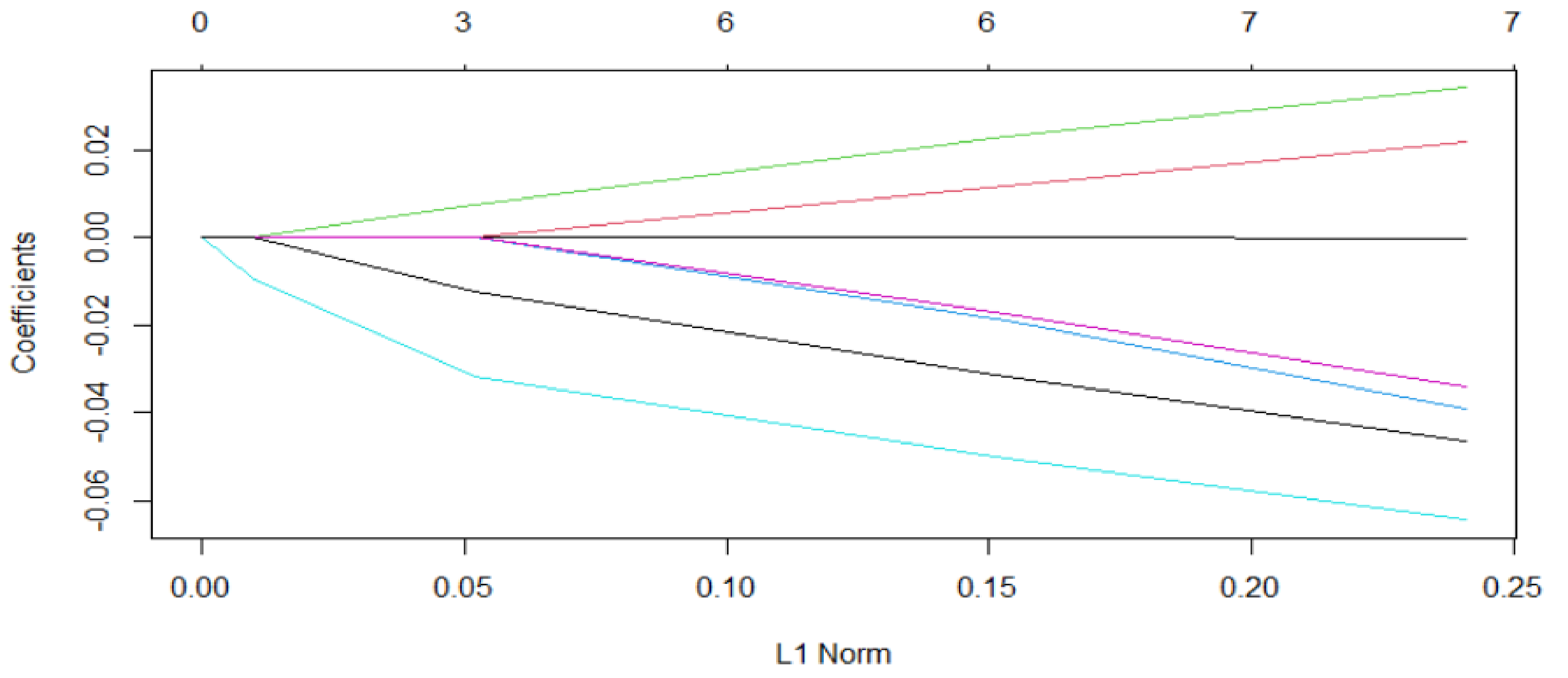


Table 3. Cross-validated Model Comparison Based on Training and Testing – Full Sample

Models	Mean Squared Error	Models	Mean Squared Error
OLS	0.09	**RIDGE**	0.11
LOGIT	0.09	**LASSO**	0.12
PROBIT	0.08	**ELASTIC NET**	0.13
LDA	0.08	**REGRESSION TREES**	0.00
QDA	0.08	**BOOSTING**	0.00

Figure 3. Cross-Validated LASSO – Standardized Lasso Coefficients in a Parsimonious Model



Conclusions

- Parsimonious models show that Primary activities in Workplace besides demographic characteristics are the best predictors.
- More complex prediction models do not outperform simpler models consistently.
- Simple Parametric models with full sample generate considerably higher MSE of prediction. Regression Boosting shows – compared to degrees, work activities are better predictors of STEM exit decision.

Contact

Md Ohiul Islam
Postdoctoral Scholar
Economics Department
University of Nevada, Reno
oislam@unr.edu

References

1. Jelks, S. M., & Crain, A. M. (2020). Sticking with STEM: Understanding STEM career persistence among STEM bachelor’s degree holders. *The Journal of Higher Education*, *91*(5), 805-831.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
3. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Statistical learning. In *An Introduction to Statistical Learning: with Applications in Python* (pp. 15-67). Cham: Springer International Publishing.
4. Luque-Fernandez, M. A., Redondo-Sánchez, D., & Maringe, C. (2019). cvauroc: Command to compute cross-validated area under the curve for ROC analysis after predictive modeling for binary outcomes. *The Stata Journal*, *19*(3), 615-625.