The predictive power of educational attainment on new firm creation: A machine learning approach

Masanori Kruoki
Professor of Economics
Arkansas Tech University
mkuroki@atu.edu

Abstract

This paper employs machine learning to investigate the predictive power of educational attainment on new firm creation across U.S. counties during the 2022-2023 period. Using county-level data on new establishment births (1-19 employee establishments) and employing various supervised learning models, we show that the inclusion of educational attainment variables improves the predictive power of linear models, increasing their out-of-sample R-squared by 0.04. We find that the share of college-educated individuals is consistently the most important predictor across all model types. While the share of high school dropouts shows lower importance in linear models, it is among the top five most important predictors in ensemble models. The findings highlight that higher educational attainment serves as an important predictive factor for establishment births at the local level, suggesting that counties with a high share of educated residents tend to experience greater creation of new firms.

Keywords: entrepreneurship; human capital; machine learning; establishment births; education

Competing Interests: None
Funding: None
Ethics and Consent to Participate declarations: not applicable

JEL code: L26, J24, R11

Introduction

One frequently studied area in the entrepreneurship literature is the role of human capital, typically measured by educational attainment. However, the empirical evidence remains somewhat mixed. The literature review by Vander Sluis et al. (2008) covering developed countries from 1980 to 2002 found no consistent evidence of education being associated with the choice to become an entrepreneur, although it did appear to increase entrepreneurs' earnings. Simoes et al. (2016) conclude that this mixed evidence stems from two conflicting effects: higher educational levels offer better job opportunities in the wage sector, yet also enhance one's ability to identify self-employment opportunities. They further suggest a potential U-shaped relationship, where entrepreneurship increases among both low-skilled and high-skilled workers. This might occur as less educated individuals transition to self-employment out of necessity (Von Greiff, 2009), while more educated individuals are better positioned to capitalize on emerging market opportunities (Bosma et al., 2004).

In U.S. studies examining regional differences, the evidence largely supports this U-shaped relationship. Acs and Armington (2006) and Armington and Acs (2002) both report a positive association between the share of high school dropouts and entrepreneurial activity, as well as between the share of college-educated adults and entrepreneurial activity. Goetz and Rupasingha (2013) found similar results for self-employment in rural areas, though they observed only the high school dropout share, not the college-educated share, associated with higher entrepreneurial activity in urban areas. Other studies (Goetz and Freshwater, 2001; Conroy and Weiler, 2017; Tsvetkova and Partridge, 2021) emphasize a strong positive influence of higher education, with college degrees often showing the largest effect on various measures of entrepreneurship.

To further advance knowledge in this subject and explore the complex statistical associations between human capital and entrepreneurship, we use machine learning methods. This approach, focusing on prediction rather than causal inference, offers a complementary perspective. Specifically, we focus on the predictive relationships between various population characteristics, with particular emphasis on educational attainment, and small business establishments. These methods allow us to assess the predictive importance of each area characteristic in predicting the outcome. If education plays a significant role in predicting local entrepreneurial activity, we should expect its importance to be clearly reflected in our prediction models.[1]

Using detailed county-level data obtained from the Business Dynamics Statistics for U.S. counties during the 2022-2023 period, our findings show that the share of college-educated individuals is consistently the most important predictor of establishment births across all model specifications. The results from our analysis indicate distinct predictive contributions of education across model types. For linear models, the inclusion of educational attainment variables leads to an approximate out-of-sample R-squared increase of 0.04. In contrast,

[1] In recent years, the use of machine learning methods has become more common among researchers (e.g., Cafarella et al., 2023; Cengiz et al., 2022; Shekhar et al., 2023).

ensemble models demonstrate more modest incremental gains in predictive power from education, with out-of-sample R-squared increases ranging from 0.02 to 0.03. Notably, while the share of high school dropouts shows relatively lower importance in linear models, its significance is considerably higher among the top predictors in ensemble models, suggesting the presence of complex, non-linear relationships that these advanced models are able to capture. These results highlight that higher educational attainment serves as a crucial predictive factor for establishment births at the local level.

Data

For our entrepreneurial activity measure, we use the number of firm entries obtained from the Business Dynamics Statistics (BDS) (U.S. Census Bureau, 2025a). Unlike other business establishment data available from the U.S. Census Bureau, such as the Nonemployer Statistics (NES) and the County Business Patterns (CBP), which provide counts of existing establishments, BDS directly tracks firm births along with detailed information on job creation by firm age and size. This makes BDS data generally preferred for directly measuring new entrepreneurship. The smallest firm size category available in BDS is 1-19 employees, which we use in this study. To facilitate comparisons of entrepreneurial density across counties of varying sizes, the number of establishment births is converted to the number of establishment births per 1,000 residents in the respective county.

We use the sample period of 2022-2023, which is the most recent period available as of October 2025. This period allows us to focus on the post-initial-shock phase of the pandemic economy, partially mitigating the immediate, disruptive influence of the 2020 lockdowns and early relief programs on firm creation while capturing the subsequent period of elevated entrepreneurial activity. Utilizing a two-year period instead of a single year also enhances the robustness of our analysis by increasing the sample size and smoothing out potential annual volatility. Counties from all U.S. states except Alaska, Hawaii, and Connecticut are included. Alaska and Hawaii are excluded because the natural amenity score, a predictor in our model, is not available for these states. All counties in Connecticut are excluded due to the state's adoption of nine planning regions as county-equivalent geographic units in 2022, which complicates the merging of various datasets. Our final sample comprises 5,874 county-year observations, representing approximately 3,000 unique counties annually.

Figure 1 displays the distribution of establishment births per 1,000 residents for firms with 1-19 employees during the 2022-2023 period. The histogram reveals a right-skewed distribution, indicating that the majority of counties exhibit lower rates of entrepreneurial activity, with a concentration of observations between 1 and 2 establishment births per 1,000 residents. The mean establishment birth rate is 1.63 per 1,000 residents, while the median is slightly lower at 1.4 per 1,000 residents, reflecting the influence of higher values pulling the mean upwards. To enhance the clarity of the distribution, a small fraction of outlier counties with extremely high birth rates are excluded from this histogram. Specifically, observations with

establishment births greater than 6 per 1,000 residents, which represent 0.5 percent of the total observations and include the maximum value of 17.98, are not shown in this figure.

[Figure 1 here]

Our main predictors of interest are education measures. Specifically, we use the share of college graduates and the share of high school dropouts, both obtained from the U.S. Census Bureau's American Community Survey (ACS) 5-year estimates (U.S. Census Bureau 2025b). As noted above, both measures have been widely used. We employ the 5-year estimates to ensure maximum coverage across counties. Figures 2 and 3 illustrate the contrasting geographic patterns of educational attainment across the United States. Figure 2 shows that high concentrations of college graduates (≥ bachelor's degree) are primarily found in major metropolitan areas and along the Northeast and West Coasts. These high-attainment areas also often correspond to counties containing large universities or government hubs, such as the Washington D.C. metropolitan area, some counties in Colorado and Florida. In contrast, Figure 3 shows that the highest proportions of dropouts are concentrated across the Southern and Appalachian United States, including Texas, the Southwest, and certain rural parts of the Midwest and West.

[Figures 2 and 3 here]

For other predictors, we include various measures of socioeconomic and demographic county characteristics. For our economic variables, we include median income and poverty rate, obtained from the U.S. Census Bureau's (2025c) Small Area Income and Poverty Estimates (SAIPE) program, and the unemployment rate, obtained from the Bureau of Labor Statistics (2025), to account for local economic conditions. Additionally, from the ACS, we obtain Gini coefficients and labor force participation rates. We also collect the share of homeowners paying more than 30% of their income for their mortgage and the share of renters paying more than 30% of their income for rent, to capture housing market conditions.

For demographic characteristics, the model incorporates the shares of Black, Hispanic, and Asian populations, as well as the share of the population aged 65 or older, and the male-to-female sex ratio. These variables are obtained from the U.S. Census Bureau's (2025d) Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin. We also include median age and the share of men and women over age 16 who are married to capture marriage market conditions. Finally, we add the share of the foreign-born population, obtained from the ACS 5-year estimates. These variables are largely chosen based on findings from previous studies that highlight the importance of age distribution (Shane, 1996), ethnic diversity (Goetz & Rupasingha, 2013), and rates and age of immigrants (Deller et al., 2019) in entrepreneurship research.

We also incorporate two time-invariant characteristics for each county. First, we include a natural amenity score, which is found to be important in past studies, such as Rappaport (2007) and Conroy and Watson (2023), that find that scenic locations with moderate climates often attract creative and educated people who value natural amenities and thereby increasing entrepreneurship in those areas. We obtain the natural amenity score produced by the U.S. Department of Agriculture's Economic Research Service (2025). This Natural Amenities Scale

data, available for counties in the lower 48 states, was constructed by combining six measures of climate, topography, and water area that reflect environmental qualities most people prefer: warm winter, winter sun, temperate summer, low summer humidity, topographic variation, and water area. The score values range from 1 to 7.

Second, we include metropolitan area classifications to account for general rural-urban differences in establishment births (Deller et al., 2019) and the differential effect of education on entrepreneurship in urban and rural areas (Goetz and Rupasingha, 2013). For this purpose, we utilize the 2023 Rural-Urban Continuum Codes obtained from the U.S. Department of Agriculture's Economic Research Service (2023). These codes distinguish U.S. metropolitan (metro) counties based on the population size of their metro area, and nonmetropolitan (nonmetro) counties by their degree of urbanization and adjacency to a metro area.

Finally, we add several state-level characteristics: state GDP per capita, state unemployment rate, state poverty rate, state minimum wages, and a dummy variable indicating whether the state governor is a Democrat. These state-level variables are obtained from the University of Kentucky Center for Poverty Research (2025). In total, we have 25 predictors. Table 1 provides summary statistics for all 25 predictors used in this study.

[Table 1 here]

Methods

Machine learning models: Linear models

In addition to standard linear regression, we also employ three other linear models: ridge, lasso, and elastic net regression. These regularized linear models are particularly valuable in a setting with many county-level predictors that are likely to be highly correlated (multicollinearity).

Ridge regression is one of the most commonly used alternatives to standard linear regression. It is also a linear model for regression, but its coefficients are shrunk towards zero. This constraint is called regularization, a technique used to restrict a model and thus avoid overfitting. Overfitting occurs when a machine learning model captures too much detail from the data it is built upon, including noise and irrelevant patterns. An alternative to ridge for regularizing linear regression is lasso. As with ridge regression, lasso also restricts coefficients to be close to zero; however, some coefficients are forced to be exactly zero with Lasso. This ability to zero out coefficients performs principled variable selection, implying that some predictors may be entirely ignored by the model. For both ridge and lasso, the balance between model simplicity (regularization strength) and goodness-of-fit on the observed data can be specified by the user using the alpha parameter. In general, a lower alpha allows a model to fit the data more closely (a more complex model), while a higher alpha forces coefficients to move more toward zero, which decreases its immediate fit to the observed data but typically improves generalization to new, unseen data. Finally, elastic net regression, which combines the penalties of both lasso and ridge regression, is also employed, for which the penalty term is a linear combination of the lasso and ridge methods. Thus, in total, four linear models are used in our analysis.

Machine learning models: Ensemble models

Ensembles are a class of powerful machine learning techniques that combine different models to create more robust and accurate predictors. They can capture complex, non-linear relationships and interaction effects between predictors, allowing us to explore the plausibility of the U-shaped relationship hypothesized in the human capital literature (e.g., between both low and high education and firm creation). In this study, we use four popular ensemble models: random forest, gradient boosting regressor, XGBoost regressor, and CatBoost regressor. These models are essentially collections of decision trees, where each tree is slightly different from the others, and their individual predictions are aggregated (e.g., averaged) to form a final result. A decision tree operates like a flowchart or a set of hierarchical "if-then-else" rules. It proceeds by making a series of decisions: at each step, the tree performs a "test" or evaluates a "condition" on a specific predictor. Based on the outcome of that test, the data follows a specific path to the next condition, eventually leading to a final prediction for the outcome variable. In general, ensemble models tend to exhibit superior predictive performance compared to simpler linear models. For illustrative purposes, Figure A1 in the appendix presents a visualization of a decision tree built using the entire dataset, with its depth limited to three levels.

A random forest constructs a multitude of fully grown decision trees, each trained on a different bootstrap sample drawn from the observed dataset. While individual trees in the forest tend to produce predictions with high variance, their collective averaging effectively reduces this variance, leading to more stable and accurate predictions. In contrast, the boosting approach (gradient boosting, XGBoost, CatBoost) also combines multiple trees, but instead of growing them simultaneously and averaging their results, trees in boosting models are grown sequentially. Each subsequent tree attempts to correct the errors of the preceding ones, iteratively improving the overall model's performance. While individual decision trees offer clear interpretability, complex ensemble models such as boosting algorithms and random forests are often referred to as "black box" models and do not readily provide insight into the specific magnitude or positive/negative direction of individual predictor associations with the outcome, unlike the coefficients found in linear models. To address this issue, we use SHAP (SHapley Additive exPlanations) values for feature importance and Partial Dependence Plots (PDPs) for visualizing marginal effect relationships, as explained below.

Measuring the importance of education

We use two metrics to measure the importance of education on establishment births. First, we utilize R-squared and assess the influence of education by comparing the change in R-squared between models with and without the education variables. However, it is important to emphasize that the R-squared values reported in this study are "out-of-sample" R-squared values. Unlike in traditional econometrics, where the primary purpose is often to measure association or examine causality, when models are used for prediction, "in-sample" R-squared is not of interest. This is because we cannot use the same data employed to build the model for its evaluation. In-sample R-squared reflects only the fit of the existing data and does not indicate whether our model will

generalize well (i.e., perform effectively on new, unseen data). To accurately assess a model's predictive performance, we evaluate it on data the model has not encountered during the model-building process. This is typically achieved by splitting the data into two parts: a training set, used to build the machine learning model, and a test set, used to assess its performance. The performance metric derived from the test set is the out-of-sample R-squared, which specifically measures how well the model predicts on new data.

To ensure a more stable and thorough evaluation of generalization performance and to avoid potential bias from a single split, we employ a technique called cross-validation. Specifically, we utilize k-fold cross-validation, a widely used statistical method where the dataset is repeatedly partitioned into subsets. In this study, we implement 5-fold cross-validation: the data is first partitioned into five equally sized subsets, or folds. Then, a sequence of five models is trained. In each iteration, one fold serves as the test set (for evaluating performance on unseen data), while the remaining four folds are used as the model-building set. The model is built on the four folds, and its performance (e.g., R-squared) is evaluated on the single held-out fold. This process is repeated five times, with each fold serving as the test set exactly once. This procedure yields a sample of five out-of-sample R-squared values, providing a robust estimate of the model's predictive performance on new data. These five R-squared values are then averaged to obtain a single, more reliable overall R-squared value for the model, which enhances the credibility and robustness of our evaluation.

One issue with using standard cross-validation for panel data is that observations for each county across different years are highly related. Allowing the same counties to appear in both the training and test sets would artificially inflate performance estimates, as the model would merely be interpolating within known county-specific patterns rather than generalizing to genuinely unseen entities. To accurately evaluate the model's generalization capability to new counties, we must therefore ensure that the training and test sets contain no overlapping counties. Thus, in our k-fold cross-validation, we employ a grouped splitting strategy where all observations belonging to the same county are kept together within either the training or the test set, preventing any single county from appearing in both simultaneously.

Our second method for quantifying the importance of education on establishment births involves two approaches: (1) comparing the magnitudes of coefficients from various linear models, and (2) obtaining predictor importances from ensemble models. We scale our predictors using standardization (removing the mean and scaling to unit variance, resulting in a mean of 0 and a standard deviation of 1) for a meaningful comparison of coefficient magnitudes across different predictors. We then calculate the average (absolute) coefficients from the four linear models to represent predictor importance. For our ensemble models, we use permutation importance (PI). This method quantifies a feature's contribution by measuring the decrease in the model's out-of-sample R-squared when the values of that feature are randomly shuffled. These importances are normalized values between 0 and 1 for each predictor, where 0 indicates no contribution to the model's predictions and higher values denote greater influence. The sum of all

predictor importances always equals 1. We calculate the average predictor importance across the four ensemble models to determine the importance of each predictor.

All machine learning models, except for standard linear regression, were tuned to optimize their hyperparameters for minimizing out-of-sample errors, employing 5-fold cross-validation. For ridge regression, the optimal alpha was found to be 0.01. For lasso regression, the optimal alpha was determined to be 0.0001. In the case of elastic net regression, the optimal alpha was 0.001, and the l1_ratio (which controls the mixing proportion of the L1 and L2 penalties) was set to 0.7. For the gradient boosting regressor, we used 200 trees and a learning rate of 0.1. The XGBoost regressor was configured with 300 trees and a learning rate of 0.05. For the CatBoost regressor, we utilized 300 trees and a learning rate of 0.1. Finally, the random forest regressor was set with 200 trees. All models were implemented using Python's scikit-learn (sklearn) library, except for CatBoost regressor (CatBoostRegressor library) and XGBoost regressor (xgb library).

To provide a more rigorous measure of variable importance for our ensemble models, we augment the standard Permutation Importance (PI) analysis with SHapley Additive exPlanations (SHAP). SHAP values are considered a superior interpretability metric, as they are derived from cooperative game theory and offer the only solution that satisfies fundamental properties of feature attribution, notably local accuracy and consistency. Finally, beyond the magnitude of importance, we use partial dependence plots (PDPs) to visualize the functional form of the relationship between specific education variables and establishment births. While linear models assume a straightforward, monotonic relationship described by a single coefficient, the power of ensemble models lies in their ability to capture complex, non-linear relationships. PDPs show the average marginal effect of a feature on the prediction, effectively illustrating the shape of the relationship while accounting for the effects of all other features. This visualization is essential for providing evidence of potentially non-monotonic effects—for instance, if the impact of the "share of college-educated" is positive up to a certain point and then declines—which would provide deeper insight than what is revealed by a single permutation importance score.

Results

Table 2 presents the out-of-sample R-squared values from eight machine learning algorithms. The linear models (linear, ridge, lasso, and elastic net regression) all show identical out-of-sample R-squared values, which increase by 0.042 when the share of college graduates and high school dropouts are included. The ensemble models (gradient boosting, XGBoost, CatBoost, and random forest) generally demonstrate higher overall out-of-sample R-squared values than the linear models. For instance, CatBoost achieves the highest R-squared of 0.585 with the education variables included. However, the increase in R-squared attributed to the inclusion of education variables is smaller for these ensemble methods, ranging from 0.022 to 0.034. This suggests that while educational attainment contributes to predictive power, its incremental benefit is less pronounced in the more complex ensemble models, which may already capture a substantial amount of the variance from other predictors.

[Table 2 here]

Figure 4 presents the relative importance of different county-level and state-level predictors for our linear models, based on standardized coefficients. This figure provides a valuable visual summary of which factors linearly contribute most strongly to the prediction of entrepreneurial activity in our dataset. Each bar reflects the absolute magnitude of a predictor's average coefficient across the linear models. Readers interested in the specific sign of each coefficient can refer to Table A1 in the appendix, which shows the full results of the linear regression model using the entire sample. The figure reveals that the share of college graduates stands out as the most influential predictor, exhibiting the largest average coefficient size. This highlights the substantial role of higher education in explaining variations in entrepreneurial activity within linear models. Conversely, the share of high school dropouts has a relatively low average coefficient size, ranking significantly lower than college graduates. As shown in Table A1 in the appendix, the negative coefficient for the share of high school dropouts suggests that a higher proportion of less-educated individuals is associated with lower new firm creation, though the coefficient is statistically insignificant.

[Figure 4 here]

Figure 5 presents the predictor importances derived from the ensemble models, showing each predictor's contribution to the models' predictive power. Consistent with the findings from linear models, the share of college graduates remains the most influential predictor of establishment births across these ensemble models. However, the ranking of other variables differs considerably from what was observed with the linear models in Figure 4. Notably, the share of the Black population and the share of high school dropouts appear among the top five most important predictors, representing a substantial increase in relative importance for high school dropouts compared to their lower ranking in the linear models. This divergence in predictor rankings highlights how different modeling approaches, particularly the non-linear relationships captured by ensemble methods, can attribute varying levels of predictive power to the same set of predictors. Nevertheless, these results consistently confirm that the college graduate share is important for predicting establishment births at the county level.

[Figure 5 here]

Figure 6 presents the SHAP-derived feature importance for all four ensemble models: gradient boosting, CatBoosting, XGBoosting, and random forest. Across all models, the results consistently confirm that the share of college graduates is the single most important predictor of establishment births across the four non-linear models. Regardless of the method used (linear vs. non-linear) or the importance ranking technique (Permutation Importance (PI) versus SHAP), the share of college graduates remains the most important predictor, which is a reassuring finding. Conversely, the importance of the share of high school dropouts is ranked lower using SHAP than with PI, generally ranking between the 10th and 13th most important predictor.

[Figure 6 here]

Finally, we use Partial Dependence Plots (PDPs) to visualize the marginal effect of educational attainment and determine if non-linear relationships are present. We select the

random forest model for this illustrative purpose because, as the simplest of our ensemble models, it provides an effective demonstration of non-linearity without the added complexity of boosting algorithms. The resulting PDPs for the two key education variables are shown in Figure 7. The panel depicting the share of college graduates reveals a positive relationship with establishment births across the entire range of predictor values, with the relationship largely linear. Conversely, the panel for the share of high school dropouts shows a generally negative relationship with establishment births. The initial effect is characterized by a sharp, steep decline in the predicted outcome as the dropout share increases below the mean. As the share of dropouts rises past the mean, the negative slope flattens considerably, suggesting the most significant decrease in predicted establishment births occurs at the lower end of this variable's distribution.

[Figure 7 here]

Conclusions

To contribute to the ongoing discussion and address the mixed evidence on the association between high rates of firm creation and a local population's average educational attainment, this study approaches the question from a predictive standpoint: is educational attainment an important predictor for new business creation? To answer this question, we depart from traditional econometric approaches and apply modern machine learning techniques. By using different types of linear regression and ensemble methods, we sought to identify the most influential predictors of new firm creation.

Our findings consistently underscore the important predictive role of higher education. Across all linear models, the inclusion of both college graduate and high school dropout shares significantly improved predictive performanc. More importantly, the share of college-educated individuals is consistently the most important predictor of establishment births, demonstrating the largest coefficient magnitudes in linear models and the highest average importance scores in ensemble models. This strong predictive association of college education with entrepreneurial activity aligns with a segment of existing literature suggesting that higher education is strongly associated with the knowledge and skills necessary to capitalize on market opportunities.

The consistent finding that higher educational attainment is a crucial predictor of local entrepreneurial activity is useful for data-driven targeting and risk assessment. The results indicate that counties with a high share of college graduates are statistically more likely to already exhibit the characteristics of a vibrant entrepreneurial ecosystem and may therefore be high-propensity areas for new firm creation. Policymakers can use these predictive insights to identify such areas for targeted support or resource allocation. For instance, identifying and supporting communities with high educational attainment and local amenities attractive to skilled workers may offer a higher likelihood of alignment with existing local conditions associated with establishment births.

References

Acs, Z. J., & Armington, C. (2006). Entrepreneurship, geography, and American economic growth. Cambridge University Press.

Armington, C., & Acs, Z. J. (2002). The determinants of regional variation in new firm formation. Regional Studies, 36(1), 33–45.

Åstebro, T., Chen, J., & Thompson, P. (2011). Stars and misfits: Self-employment and labor market frictions. Management Science, 57, 1999–2017.

Bosma, N., Van Praag, M., Thurik, R., & De Wit, G. (2004). The value of human and social capital investments for the business performance of startups. Small Business Economics, 23, 227–236.

Cafarella, M., Carriero, G., Clark, T., & Marcellino, M. (2023). Using Machine Learning to Correct for Quality Change When Measuring Inflation. NBER Working Paper, No. 31393.

Cengiz, D., Dube, A., Lindner, A., & Zentler-Munro, D. (2022). Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes. Journal of Labor Economics, 40(S1), S203–S247.

Conroy, T., & Watson, P. (2023). Overeducation, natural amenities, and entrepreneurship. Small Business Economics, 61, 1111–1131.

Conroy, T., & Weiler, S. (2017). Where are the women entrepreneurs? Business ownership growth by gender across the American urban landscape. Economic Inquiry, 53(4), 1872–1892.

Deller, S., Kures, M., & Conroy, T. (2019). Rural entrepreneurship and migration. Journal of Rural Studies, 66, 30–42.

Goetz, S. J., & Freshwater, D. (2001). State-level determinants of entrepreneurship and a preliminary measure of entrepreneurial climate. Economic Development Quarterly, 15(1), 58–70.

Goetz, S. J., & Rupasingha, A. (2013). The determinants of self-employment growth: Insights from county-level data, 2000-2009. Economic Development Quarterly, 28(1), 42–60.

Joona, P., & Wadensjö, E. (2013). The best and the brightest or the least successful? Self-employment entry among male wage-earners in Sweden. Small Business Economics, 40, 155–172.

Poschke, M. (2013). Who becomes an entrepreneur? Labor market prospects and occupational choice. Journal of Economic Dynamics & Control, 37, 693–710.

Rappaport, J. (2007). Moving to nice weather. Regional Science and Urban Economics, 37, 375–398.

Shane, S. (1996). Explaining variation in rates of entrepreneurship in the United States: 1899–1988. Journal of Management, 22, 747–781.

Shekhar, V., Tabarrok, A., & Wager, S. (2023). Detecting Fraud and Overbilling Among US Hospitals Using Unsupervised Machine Learning. NBER Working Paper, No. 31327.

Simoes, N., Crespo, N., & Moreira, S. B. (2016). Individual determinants of self-employment entry: What do we really know? Journal of Economic Surveys, 30(4), 783–806.

Tsvetkova, A., & Partridge, M. (2021). Knowledge-based service economy and firm entry: An alternative to the knowledge spillover theory of entrepreneurship. Small Business Economics, 56(2), 637–657.

University of Kentucky Center for Poverty Research. (2025). UKCPR national welfare data, 1980-2023. Lexington, KY. Retrieved from http://ukcpr.org/resources/national-welfare-data

U.S. Bureau of Labor Statistics. (2025). *Local Area Unemployment Statistics (LAUS)*. Retrieved May 10, 2025, from https://www.bls.gov/lau/tables.htm#cntyaa

U.S. Census Bureau. (2025a). *Business Dynamics Statistics datasets*. Retrieved May 10, 2025, from https://www.census.gov/data/datasets/time-series/econ/bds/bds-datasets.html

U.S. Census Bureau (2025b). American Community Survey 5-Year Data. Retrieved May 10, 2025, from https://www.census.gov/data/developers/data-sets/acs-5year.html

U.S. Census Bureau. (2025c). *Small Area Income and Poverty Estimates (SAIPE): 2023 State and County*. Retrieved May 10, 2025, from https://www.census.gov/data/datasets/2023/demo/saipe/2023-state-and-county.html

U.S. Census Bureau (2025d) Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin. Retrieved May 10, 2025, from https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-detail.html

U.S. Department of Agriculture, Economic Research Service. (2025). Natural amenities scale. Retrieved from https://www.ers.usda.gov/data-products/natural-amenities-scale

U.S. Department of Agriculture, Economic Research Service. (2023). Rural-Urban Continuum Codes. Retrieved from https://www.ers.usda.gov/data-products/rural-urban-continuum-codes

Van Der Sluis, J., Van Praag, M., & Vijverberg, W. (2008). Education and entrepreneurship selection and performance: A review of the empirical literature. Journal of Economic Surveys, 22(5), 795–841.

Von Greiff, J. (2009). Displacement and self-employment entry. Labour Economics, 16, 556–565.

Table 1. Summary statistics

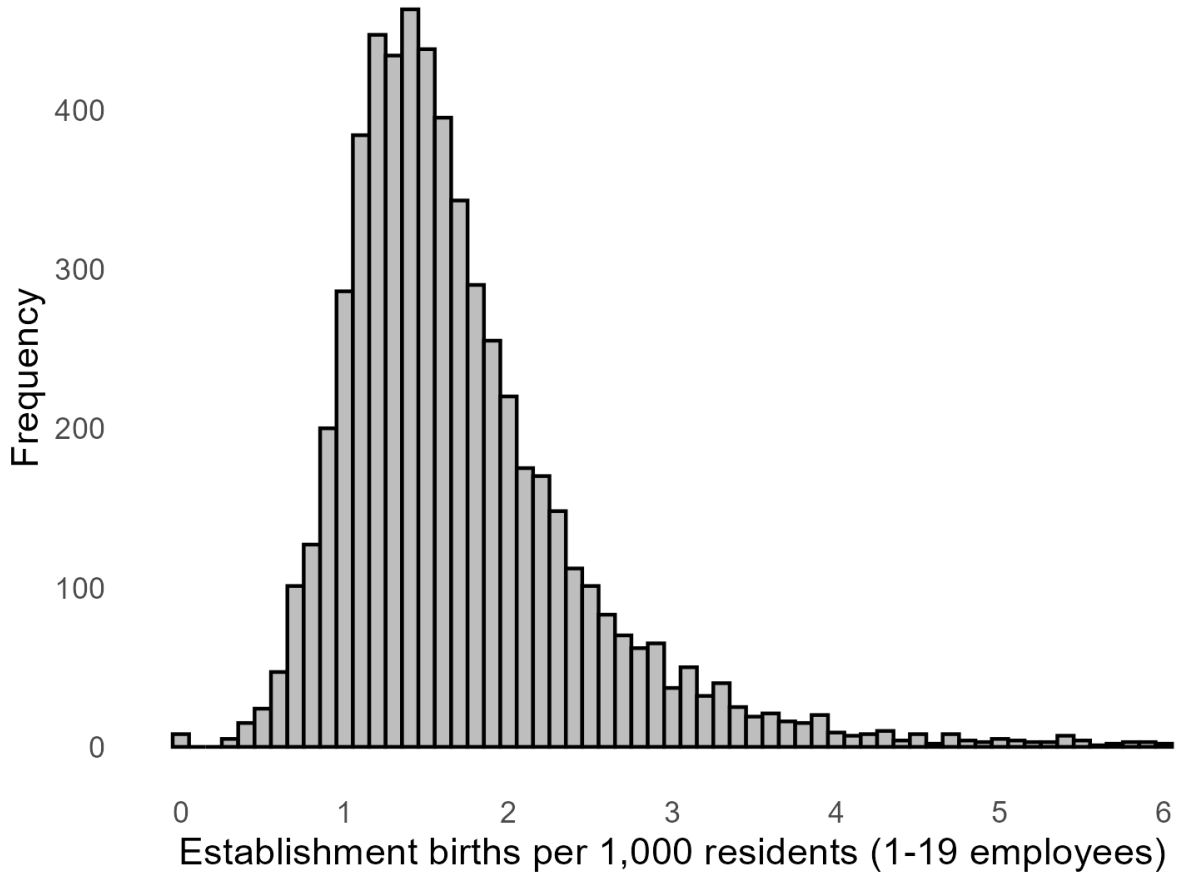
Variables	Mean	Min	Max	SD
Share College Graduates	23.92	4.4	79.73	10.18
Share High School Dropouts	11.41	0	42.94	5.42
Local Unemployment Rate	3.6	1.3	17.6	1.14
Local Poverty Rate	14.43	3.1	49.6	5.43
Median Household Income	64637.14	28579	173655	16384.98
Gini Coefficient	0.45	0.33	0.62	0.04
Homeowners with High Housing Cost	18.48	1.64	49.13	4.63
Renters with High Housing Cost	37.59	1.25	65.26	9.14
Labor Force Participation Rate	58.02	18.8	78.63	7.41
Metropolitan Area Dummy	0.39	0	1	0.49
Amenity Score	3.49	1	7	1.05
Share Black Population	9.29	0.1	85.58	14.17
Share Asian Population	1.6	0.06	41.62	2.72
Share Hispanic Population	10.29	0.28	97.17	13.61
Share Population Over 65	20.51	5.56	57.37	4.74
Median Age	41.59	21.4	68.4	5.21
Share Foreign-Born	4.78	0	47.57	5.6
Sex Ratio (Male-to-Female)	1.02	0.73	2.47	0.1
Share Men Married	51.15	19.44	77.94	6.94
Share Women Married	49.64	20.54	83.7	7.15
State Unemployment Rate	3.33	1.9	5.2	0.66
State GDP Per Capita	72831.18	47628.39	109970.78	12103.19
State Poverty Rate	11.37	6	21.5	2.98
Democratic Governor (State)	0.39	0	1	0.49
State Minimum Wage	9.26	7.25	15.74	2.45

N = 5,874

Table 2. Out-of-sample R^2 with and without education variables

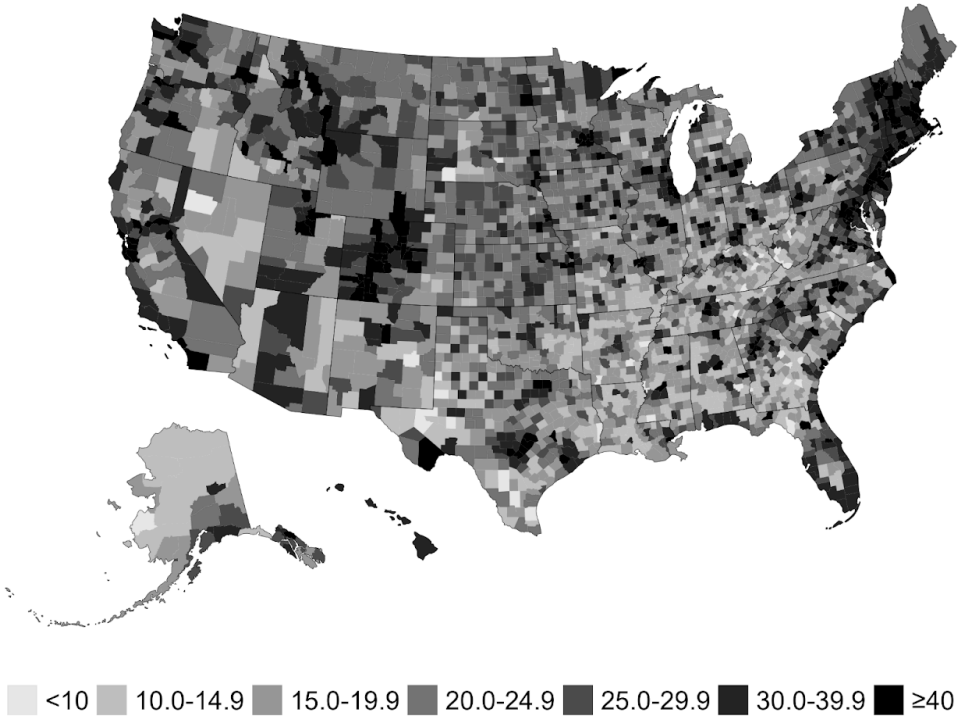
Model	R^2 without education variables	R^2 with education variables	Difference
Linear Regression	0.435	0.477	0.042
Ridge Regression	0.435	0.477	0.042
Lasso Regression	0.435	0.477	0.042
Elastic Net Regression	0.435	0.477	0.042
Gradient Boosting	0.534	0.559	0.025
XGB Boosting	0.553	0.576	0.022
CatBoosting	0.551	0.585	0.034
Random Forest	0.512	0.545	0.034

Figure 1. Distribution of county-level establishment births (1-19 employees): 2022-2023



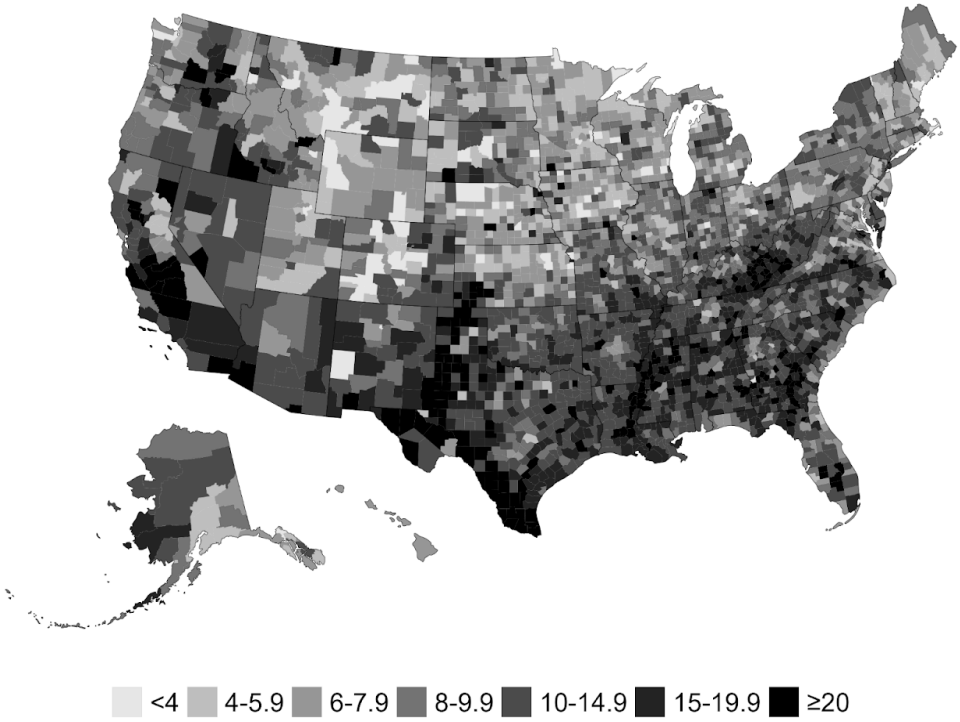
Note. Restricted to establishment births less than 6.

Figure 2. Population share of college graduates



Source: American Community Surveys 2017-2022 and 2018-2023.

Figure 3. Population share of high school dropouts



Source: American Community Surveys 2017-2022 and 2018-2023.

Figure 4. Average coefficients from linear models

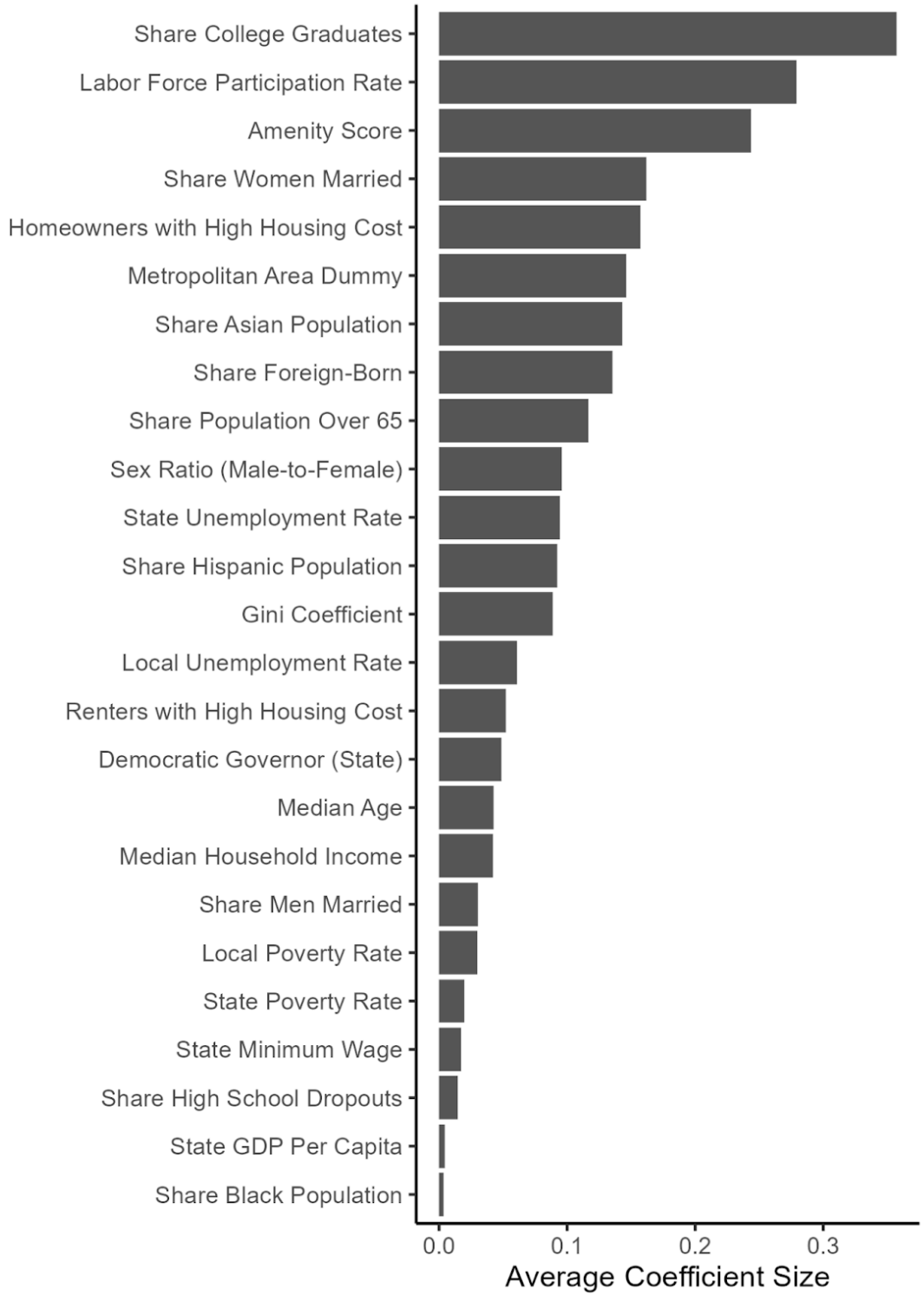


Figure 5. Average predictor importance from ensemble models

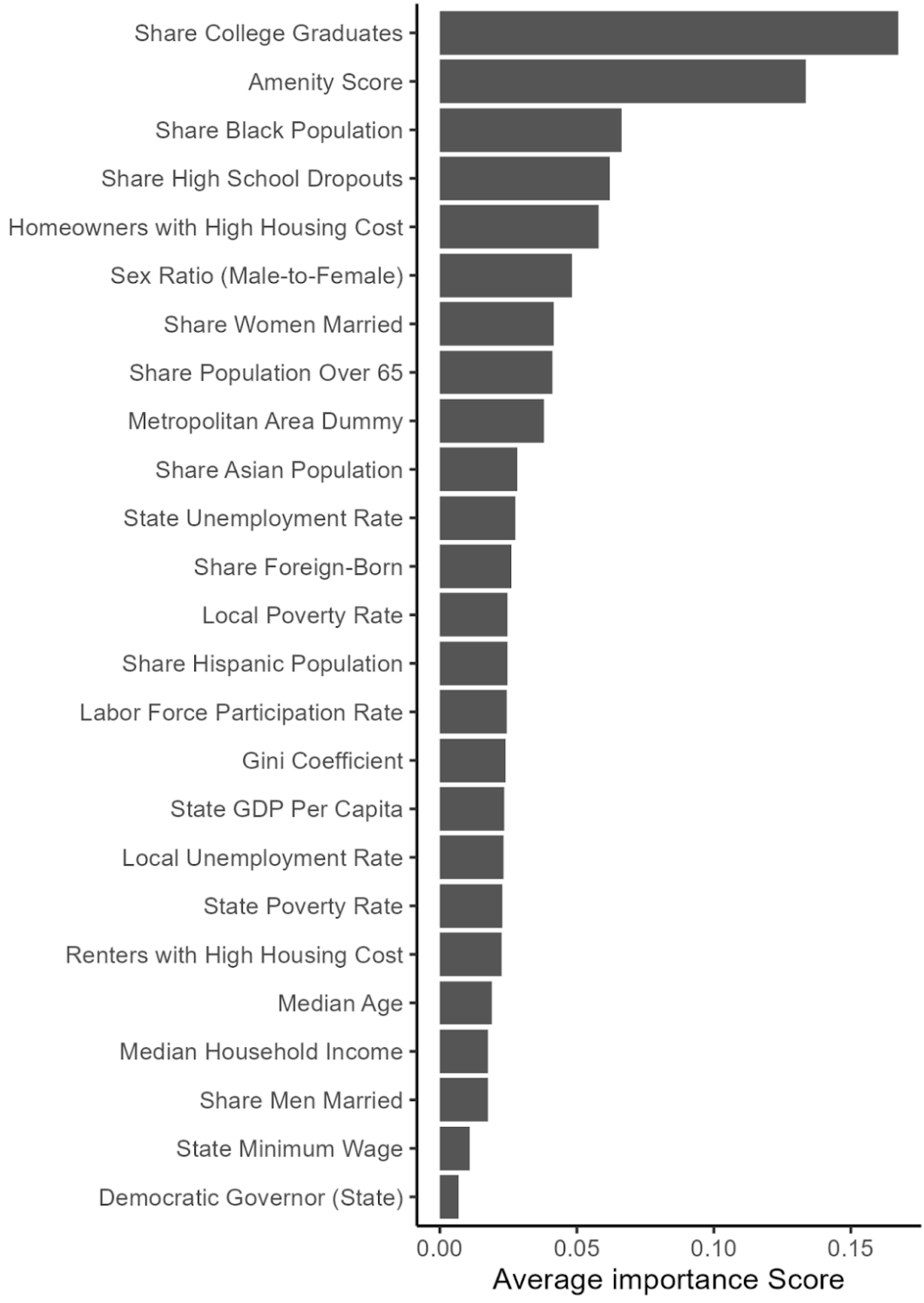


Figure 6. SHAP

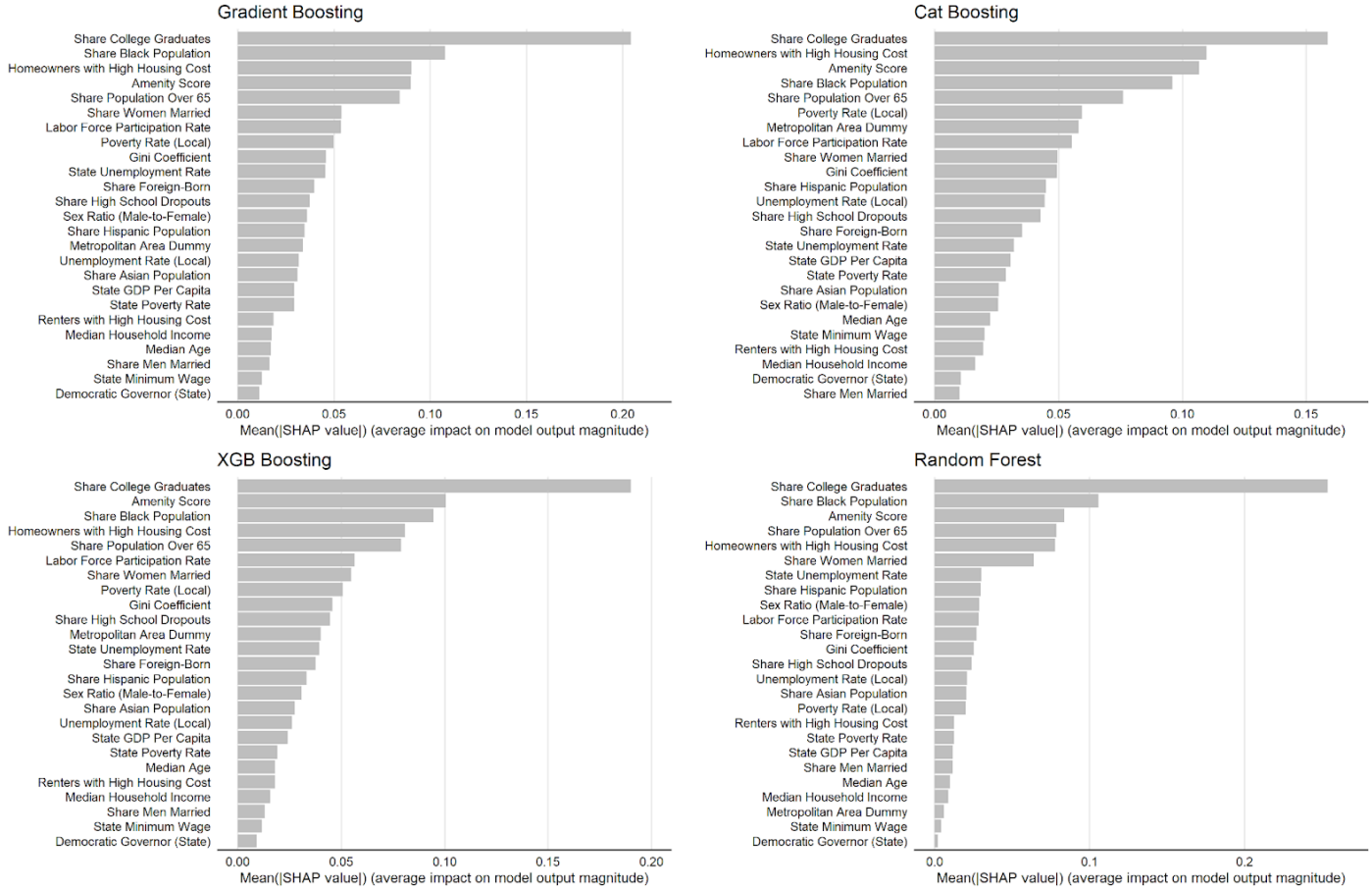
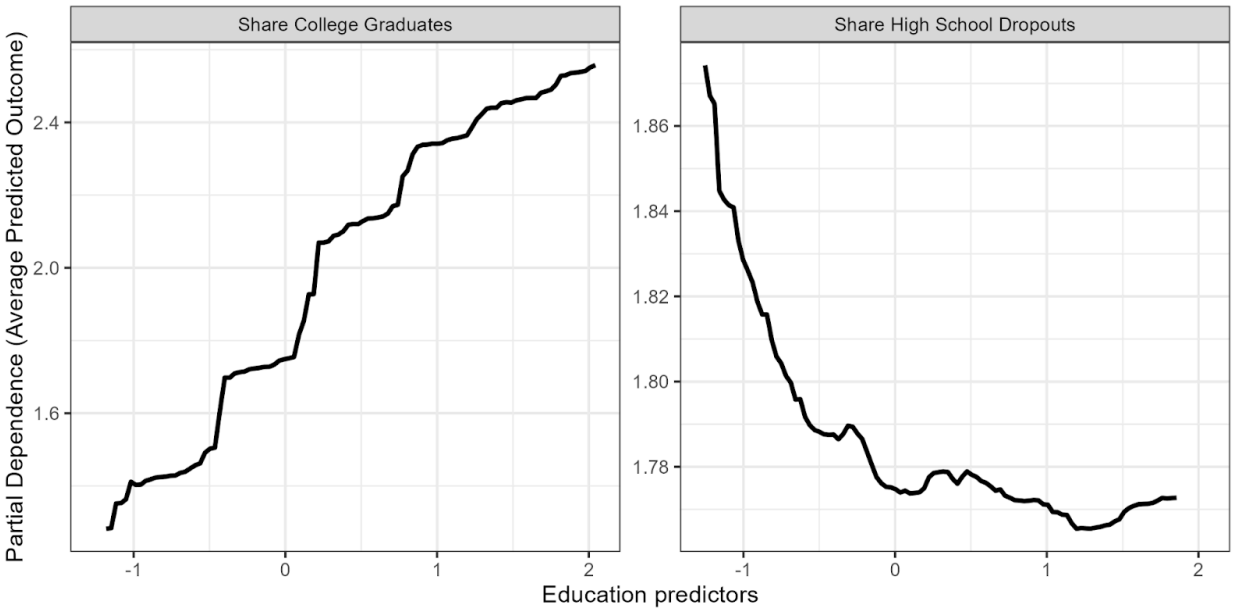


Figure 7. Partial dependence plots for random forest



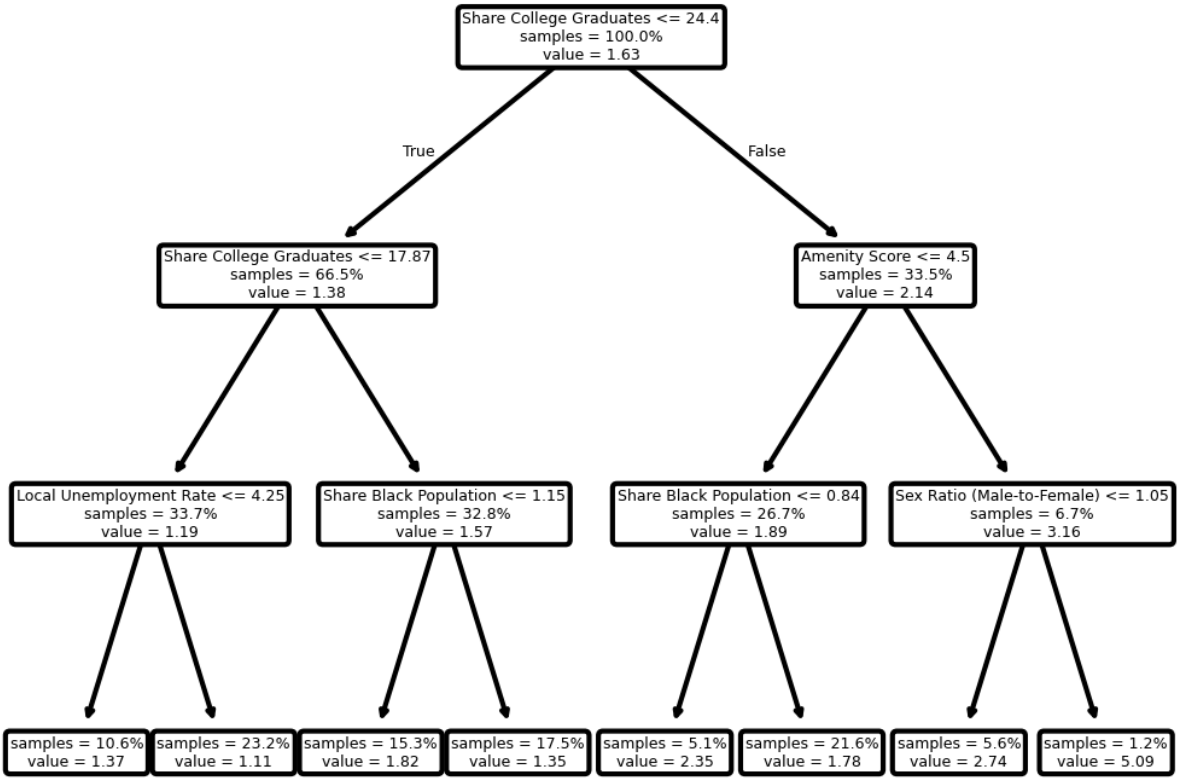
Appendix

Table A1. Linear regression results: Establishment births per 1,000 residents

Share College Graduates	0.349*** (0.018)
Share High School Dropouts	-0.013 (0.018)
Local Unemployment Rate	0.082*** (0.016)
Local Poverty Rate	-0.039* (0.023)
Median Household Income	-0.027 (0.023)
Gini Coefficient	0.099*** (0.012)
Homeowners with High Housing Cost	0.138*** (0.011)
Renters with High Housing Cost	-0.074*** (0.011)
Labor Force Participation Rate	0.294*** (0.020)
Metropolitan Area Dummy	-0.148*** (0.011)
Amenity Score	0.248*** (0.011)
Share Black Population	0.002 (0.013)
Share Asian Population	-0.152*** (0.017)
Share Hispanic Population	-0.081*** (0.016)
Share Population Over 65	0.127*** (0.027)
Median Age	0.044* (0.025)
Share Foreign-Born	0.127*** (0.021)
Sex Ratio (Male-to-Female)	0.084*** (0.014)
Share Men Married	-0.059*** (0.019)
Share Women Married	0.173*** (0.021)
State Unemployment Rate	-0.104*** (0.014)
State GDP Per Capita	0.008 (0.013)
State Poverty Rate	0.027* (0.014)
Democratic Governor (State)	-0.044*** (0.010)
State Minimum Wage	-0.017 (0.011)
Constant	1.740*** (0.009)
Observations	5,874
Adjusted R^2	0.461

*p<0.1; **p<0.05; ***p<0.01

Figure A1. Illustrative decision tree results



Note: For illustrative purposes, the tree depth is limited to three levels.