**Predicting postsecondary credential possession using the Adult Training and Education Survey of the 2016 National Household Education Survey**

Charles H. Rosemond

City University of New York School of Professional Studies

DATA 698 - Capstone

Arthur J. O’Connor

December 13, 2021

**Abstract**

State of Illinois stakeholders seek to increase postsecondary credential attainment among adult Illinoisans to meet employer demand and strengthen the state’s economy. This analysis applies classification methods to data from the Adult Training and Education Survey (ATES) component of the 2016 National Household Education Survey (NHES) to predict whether an individual possesses a postsecondary certification, certificate, or academic degree, and what characteristics inform those predictions. The predictive performance of the analysis’ models is mixed due to limitations of the ATES data as well as real imbalances in possession of certain credentials in the United States. The interpretation of important model features is thus limited, though characteristics like an individual’s earnings over the past year or their [lack of] credential area of focus appear particularly important in predicting credential possession. Future analyses should employ additional methods, focus especially on certification and certificate holders, and incorporate Illinois-sourced data to better inform education-and-workforce-system planning and decision-making.

**Introduction**

In an ideal world, all students would graduate from high school prepared for whatever comes next, whether starting a job, enrolling in postsecondary education, or enlisting in the military. The reality of the job market, however, suggests that high school graduates should pursue further education or training to ensure they are positioned to have options and pursue future opportunities. In Illinois, recent data shows that approximately 80 percent of surveyed Illinois employers seek employees with at least some postsecondary education or training (Illinois 60 by 25 Network, 2021). This education or training often culminates in a credential, like an academic degree, that serves as a signal to employers of a prospective employee’s competency or skill to perform the job.

In response, roughly a decade ago, the State of Illinois established a goal of ensuring that 60 percent of all adults in the state possess a high-quality postsecondary credential by 2025 (Illinois 60 by 25 Network, 2021). Progress towards this goal has been slow but steady, though reaching the 60 percent threshold in the next several years is increasingly unlikely. In its *A Stronger Nation* report describing credential attainment nationwide and in individual states, the Lumina Foundation finds that as of 2019, Illinois sat at approximately 55.20 percent overall, inclusive of various types of credentials (The Lumina Foundation, 2021).

Per the federal Workforce Innovation and Opportunity Act, the State of Illinois defines a credential as “a verification of qualification or competence issued to an individual by a third party with the relevant authority or jurisdiction to issue such credentials” (Illinois Department of Commerce and Economic Opportunity, 2021, “Credentials” glossary term). Different types of include educational diplomas, certificates, and degrees; registered apprenticeship certificates; occupational licenses; and professional personnel or skill-based certifications. Postsecondary academic degrees tend to differ by the level and length of the program of study, e.g., a two-year long program for an associate degree versus a four-year program for a bachelor’s degree. Similarly, certificates “require completion of an organized program of study at the postsecondary level”, though not necessarily from an academic institution, and “are classified by the amount of time required to complete the program of study” (Illinois Department of Commerce and Economic Opportunity, 2021, “Certificate” glossary term).

Illinois stakeholders across education and the workforce continue working to increase postsecondary credential attainment. Substantial recent efforts have focused on the creation and implementation of career pathways, defined as “a combination of rigorous and high-quality education, training, and other services that aligns both vertically and horizontally across secondary education, adult education, workforce training and development, career and technical education, and postsecondary education systems, pathways, and programs” (State of Illinois, 2018, p. 8). In Illinois, the pathways systems stem from the Postsecondary and Workforce Readiness (PWR) Act. Passed and signed in 2016, the PWR Act aligns school districts, postsecondary institutions, employers, and community organizations behind a competency-based approach to preparing students for their next steps after high school (State of Illinois, 2018, p. 4). Early qualitative evidence suggests that the PWR Act and related policies—including pathways endorsements on high school diplomas and frameworks organizing college-and-career advising—combine to lay a strong preparatory foundation for students.

There is limited understanding in Illinois of what characteristics predict an individual’s possessing certain types of credentials. Though the state agencies charged with administering pathways programming per the PWR Act have improved their data collections, the relative recency of such collections limits the availability of longitudinal data and quantitative data in general. Further, data element definitions are not always consistent across agency data systems. Even where the relevant data exist and are definitionally aligned, integrating them across agencies is typically inefficient. Illinois policymakers, researchers, and other stakeholders have thus not been well equipped to assess policies or better understand individual credential attainment. There is interest, however, in analyzing data from elsewhere that may generalize to the Illinois context.

To support efforts to increase postsecondary credential attainment in Illinois, this analysis will apply classification methods to publicly available, nationally representative NHES microdata to build models predicting an individual’s possession of different types of postsecondary credentials. These models could inform the planning and implementation of career pathways and related education-and-workforce-training systems in Illinois and elsewhere.

**Literature review**

A dearth of research uses data describing United States-residing adults—and, by extension, characteristics related to employment—to predict postsecondary credential attainment. However, credentials can serve as a proxy for postsecondary educational attainment, and a rich body of literature investigates attainment and its relations with general demographic, attitudinal, and other individual characteristics present among school-aged youth. This body has driven the general acceptance that postsecondary educational attainment positively correlates with socioeconomic well-being in this country.

Much historical work has sought to better understand factors predictive of postsecondary educational attainment, including race or ethnicity—the focus of equity targets in credentialing for the State of Illinois. Kao and Thompson offer a meta-analysis of various empirical research that investigates differences in educational attainment across racial/ethnic lines (2003, pp. 417-419). Examples of such research include articles from Wolfle (1985) and Thompson, Gorin, Obeidat, and Chen (2006) that compare attainment rates, and the factors that predict those rates, between Black and White individuals. The former, using data drawn from the National Longitudinal Study of the high school graduating class of 1972, finds that effects on attainment of social background—as operationalized by factors like parental occupation and education—and student characteristics like ability are similar across groups (Wolftle, 1985, p. 502). Thompson, et al. analyze somewhat fresher data, from 1992, and find that the predictive effects on postsecondary educational attainment of student gender, socio-economic status, and expectations, among other factors, vary within and across groups (2006, pp. 550-552). These patterns suggest the presence of interactions between factors.

Acknowledging such interactions as well as expanding to include more familial and environmental factors, works like Wilson and Allen (1987) and Garcia and Bayer (2005) focus on within-group variation in postsecondary education attainment. Using various longitudinal data sets from the late 1970s, Wilson and Allen identify family structure and encouragement and school encouragement, in addition to demographic characteristics, as predicting attainment among Black youth (1987, p. 71). Garcia and Bayer also investigate various family- and school-related predictors of attainment, focusing on Latino individuals and variation across ethnic groups (2005, pp. 513-522). Regarding other environmental factors, Rumbaut identifies several “turning point” events, including incarceration for males and motherhood for females, that can affect educational attainment and general opportunity among the children of immigrants to the United States (2005, p. 1041). These events and their effects are consistent across countries of origin, or ethnic groups (Rumbaut, 2005, p. 1041).

Methodologically, the literature on postsecondary educational attainment tends to feature flavors of linear and logistic regression. Using regression makes sense given the general, predominant interest in explanation over prediction. These models, as in Wilson and Allen (1987), regress dependent variables like educational attainment, ordinal or binary, on vectors of potential explanatory variables (pp. 67-68). Where appropriate, as in Thompson, et al. (2006), they incorporate interactions between variables (p. 557). Also common, as in Garner and Raudenbush (1991), is hierarchical nesting within schools or other units (pp. 252-253). There appears to be potential for applying more advanced models for predicting attainment.

Relative to postsecondary educational attainment as measured via academic degrees, research and data describing factors predicting possession of postsecondary non-degree credentials are limited. Poor definitional alignment of what constitutes a certificate or certification has hindered efforts, as have the dynamic and ever-changing landscape of non-degree credentialing programs and the limited availability of data describing that landscape. In a 2018 scan for the Workforce Data Quality Campaign, Leventoff highlights the challenges faced by states nationwide, including Illinois, in defining and collecting these data. They cite ongoing improvements in these areas across states, however (Leventoff, 2018, pp. 3-4).

In Illinois, with limited access to non-degree credential data for residents, stakeholders have sought extant evidence regardless of whether that evidence applies directly to the state. This analysis cannot address definitional or market-related challenges, nor is its proposed data set Illinois-centric. It could, however, offer a useful data point that promotes better understanding of the predictive factors driving individual adults’ interactions with the postsecondary non-degree credential landscape.

**Research questions**

This analysis focuses on using individual characteristics to predict possession of a postsecondary credential. Types of postsecondary credentials of interest include non-degree certificates or certifications and academic degrees, which include associate, bachelor’s, and professional or graduate degrees. Data definitions for the variables representing possession of these credentials are provided in *Data and Methods*.

A small set of research questions guides analysis, focusing on predicting possession of different types of postsecondary credentials. These predictions, and an understanding of the characteristics driving them, could inform building and implementing better education-and-workforce-training systems in Illinois and elsewhere.

* What characteristics predict possession of a postsecondary certification?
* What characteristics predict possession of a postsecondary certificate?
* What characteristics predict possession of a postsecondary academic degree?

The literature suggests that various demographic characteristics share strong relationships with postsecondary educational attainment, proxied here by postsecondary credential possession. Expected key predictive factors of postsecondary credential possession include race/ethnicity and gender. Employment-related characteristics like current job status, current wage, and current industry and/or occupation are also expected to have effects. Anecdotal evidence from Illinois suggests that demography plays a substantial role in credential possession, as it does in access to education and training and in interactions with other societal systems. Likewise, there are known relationships between an individual’s employment—including their wage, industry, and occupation—and the credential(s) necessary to obtain and maintain that employment.

**Data and methods**

This analysis uses data from the ATES component from the 2016 administration of the NHES, a survey program administered by the National Center for Education Statistics at the U.S. Department of Education’s Institute for Education Sciences (The National Center for Education Statistics, 2021). Focused on topics including early childhood care and education, family or parental involvement in schools, and adult education, the NHES collects microdata describing individuals living in households across the country. NHES administrations occur every three to four years using nationally representative cross-sectional samples (The National Center for Education Statistics, 2021). The 2016 NHES administration was the first and only to date to include the ATES, which targets non-institutionalized adults ages 16 to 65 who are no longer in enrolled in secondary education (The National Center for Education Statistics, 2018b, p. 1). Specific information collected by the 2016 administration included “educational attainment; the prevalence and characteristics of certifications and licenses; the prevalence and characteristics of educational certificates; and completion and key characteristics of work experience programs, such as apprenticeships and internships” (The National Center for Education Statistics, 2018b, p. 6).

Sampling for the 2016 NHES administration consisted of two phases and was stratified by both the racial/ethnic and poverty-level compositions of census tracts (The National Center for Education Statistics, 2018b, p. 8). The first phase drew from a frame of over 200,000 households, identified by street addresses, from all 50 states and the District of Columbia (The National Center for Education Statistics, 2018b, p. 8). The frame was randomly sampled, with each sampled household received a screener survey, which sought demographic information on the households to inform eligibility for the ATES and other topical surveys (The National Center for Education Statistics, 2018b, p. 8). The second phase sampled individuals, one possible per household, who were screened as eligible in the prior phase, with selection also based upon the proportion Black or Hispanic of an individual’s census tract of residence and that tract’s proportion of households in poverty (The National Center for Education Statistics, 2018b, p. 9). Specific strata for race/ethnicity included census tracts with proportion Black of at least 25 percent, tracts with proportion Hispanic of at least 40 percent, and all other tracts; specific sub-strata for poverty included census tracts with proportion below the federal poverty line of at least 20 percent, and all other tracts (The National Center for Education Statistics, 2018b, p. 9).

A randomly sampled subset of the broader ATES public-use dataset, the dataset used for this analysis contains approximately 24,000 observations, each representing an adult individual who completed the survey, and over 350 variables. Over half of the variables are frequency weights or imputation flags, which are ignored for this analysis and its predictive modeling, and many others represent responses to branched survey questions or are otherwise irrelevant to the research questions of interest. The remaining 30-odd categorical features range from the dependent variables to individual demography to employment behaviors to environmental characteristics. Following one-hot encoding of this set of categorical features, about 120 dummy features exist in the data set; this total includes the encoded dependent variables not serving as the target for a specific research question.

The dependent variables of interest—indicators of credential possession—are derived from existing variables within the dataset. Variables indicating possession of a certification (*CERTN*), a postsecondary certificate (*PSCERT*), or a postsecondary degree (*PSDEGREE*) are all binary. Table 1 describes these variables’ respective data definitions.

**Table 1**

*Data definitions of the dependent variables for analysis*

|  |  |
| --- | --- |
| Dependent variable | Data definition |
| Postsecondary certification (*CERTN*) | A postsecondary credential that was not a certificate or degree and was not required by a government agency  df$CERTN <- if\_else((df$CNMAIN == 1 &  (df$CNPROV1 %in% c(2,3) | df$CNPROV2 %in% c(2,3) | df$CNPROV3 %in% c(2,3)) &  (df$CNINVALID1 != 1 & df$CNINVALID2 != 1 & df$CNINVALID3 != 1)), 'Yes', 'No') |
| Postsecondary certificate (*PSCERT*) | A subbaccalaureate certificate earned from a postsecondary institution, that did not involve at least forty hours of instruction, and that did not require concurrent pursuit or completion of a postsecondary degree  df$PSCERT <- if\_else((df$CERTPROG == 1 &  df$LASTPSCER == 1 &  df$LCHOURS %in% c(1,2,3,4) &  df$LCENROLL %in% c(3,4)), 'Yes', 'No') |
| Postsecondary degree (*PSDEGREE*) | Possession of at least one of the following: an associate degree, a bachelor’s degree, or a graduate or professional degree  df$PSDEGREE <- if\_else((df$EDUC %in%  c('AssociateDegree',  'BachelorsDegree',  'GradDegree')),'Yes','No') |

*Note:* The National Center for Education Statistics (2018a). *Derived variables used in the NHES Adult Training and Education Survey (ATES) First Look Report (NCES 2018-103rev)*.

Two of the dependent variables—*CERTN*, at approximately seven-percent possession, and *PSCERT*, at approximately ten-percent possession—are highly imbalanced between classes. Given that the ATES data are nationally representative, it is reasonable to assume that they reflect the true population proportions of certification and postsecondary certificate possession, respectively. To artificially balance these variables during model training, whether through under- or oversampling or a combination, could thus introduce bias. In response, simple class weights reflecting the ratio between majority and minority classes, e.g., the number of individuals possessing a certification versus the number not, are used to attempt to address the imbalance. Additionally, models predicting *CERTN* and *PSCERT* are evaluated using the area under the receiver operating characteristic curve, or AUC-ROC, rather than accuracy, which is not robust to imbalanced classes.

Feature selection employs a series of chi-square (χ2) tests of independence between the dependent variable(s) and each of the possible features. Most features share statistically significant relationships (α = 0.05) with each dependent variable per the χ2 test statistic. Of the three dependent variables, *PSDEGREE* generally returns the largest test statistics. Additional feature selection comes via checks for near-zero variance, which, if present, indicates that all or nearly all observations share the same value for a given feature. Roughly a quarter of features display near-zero variance and are thus removed prior to modeling. Post-feature selection, between roughly 75 and 85 features are available for modeling towards each research question, respectively.

Missing data are minimal in the ATES variables used for this analysis. Imputing meaning for missing values—meaning where it may not exist—can be problematic, particularly with limited domain expertise. Given the ATES data are derived from responses to survey questions with frequent branch logic, they contain substantial but valid missing data, or valid skips, for features associated with questions down-branch. This analysis is primarily interested in the initial indicator questions that prompt branching and thus avoids these valid skips. Otherwise, the data set is essentially complete; a negligible number of rows including non-skip-related missing values are dropped prior to modeling.

The research questions call for the use of classification to predict possession of different types of postsecondary credentials. For each question, the dependent variables of interest, all labeled, warrant a comparison of different supervised classifiers. The three specific methods are described below at a high level along with the relevant model parameters tuned for this analysis.

The random forests algorithm is an ensemble method that introduces randomness to the construction of tree learners in a bagged tree model. This randomness, done via randomly selecting predictors at each split of each tree in the ensemble, represents an attempt to limit correlation between predictors, reduce variance of the overall ensemble, and thus limit variance in subsequent predictions. Here, a single tuning parameter is used: the number of predictors, or *m*try, chosen at each split (Kuhn and Johnson, 2016, p. 199-200).

The gradient boosting machines (GBM) algorithm is, like random forests, an ensemble method that uses trees as learners. However, unlike the independent trees of the random forests, it builds an additive model of dependent trees that greedily selects the optimal tree at each stage of the algorithm. Shrinkage is used to reduce any resulting over-fitting. Here, several tuning parameters are used: the shrinkage rate, or the proportion of each iteration’s predicted value that is added to the prior one; tree depth, or the depth of each tree in the ensemble; the number of trees in the ensemble; and the minimum number of observations to be included in the terminal node of each tree in the ensemble (Kuhn and Johnson, 2016, p. 205-206).

The extreme gradient boosting algorithm, or XGBoost, extends gradient boosting, enabling parallel computation that results in substantial time savings (xgboost developers, 2021). Here, several tuning parameters are used: the number of rounds, or passes on the data; the max depth of each tree in the ensemble; eta (η), or the shrinkage rate; gamma (γ) or the minimum loss reduction necessary to extend a tree node; the number of predictors, or columns, chosen when constructing each tree; and the subsample of data that the algorithm uses to grow its trees (Srivastava, 2016).

All models are trained on a training set consisting of approximately 60 percent of observations from the broader data set, or roughly 14,300 observations. The trained models are each cross-validated using a ten-fold scheme. They are then evaluated and compared via predictive performance on a test set representing 20 percent of observations, or roughly 4,700 observations. The model showing the highest performance on the test set is then validated on a validation set representing the remaining 20 percent, or roughly 4,700 observations.

**Findings and discussion**

For each research question, random forests, GBM, and XGBoost classifiers are trained on a training set—finding the optimal tuning of the parameters noted in *Data and Methods—*and evaluated on a test set. Evaluation metrics include AUC-ROC for the imbalanced dependent variables *CERTN* and *PSCERT* and accuracy for *PSDEGREE*. The highest performing model per these metrics is then validated on a final validation set and assessed to determine the features most important in informing its predictions.

***What characteristics predict possession of a postsecondary certification?***

Modeling of *CERTN* begins with a series of non-weighted models to establish how the algorithms react to the imbalance between classes. In the ATES dataset and more broadly in the United States, a small proportion of the adult population possesses a postsecondary certification. Regardless of the relatively large number of observations and the dozens of features available, capturing and predicting individuals making up this proportion could prove challenging. And such is the case for this analysis.

Per accuracy, the non-weighted random forests, GBM, and XGBoost models predict *CERTN* in the test set well. However, this metric is not robust to the imbalanced classes in the dependent variable and thus suggests strong performance where none exists. Reacting to the approximately seven percent of observations indicating possession of a certification, post-training, the optimally tuned versions of each model predict that zero test-set observations possess a certification. That is, each model simply predicts that all observations are a “no”, resulting in an accuracy of roughly 93 percent for each one. The remaining seven percent of observations—representing those individuals possessing a certification—are false negatives. Such predictive behavior is reasonable but not ideal.

Two actions attempt to address the issue in a second round of modeling. First, case weights representing the ratio of credential non-holders to holders—roughly 13.80 applied to holders and 1.00 applied to non-holders—are applied during training of each model. Second, AUC-ROC is used as the primary evaluation metric, alongside sensitivity—essentially, the model’s ability to predict true credential holders—and specificity—the model’s ability to predict true credential non-holders. Table 2 contains these metrics, rounded to three decimal places, for each weighted model.

**Table 2**

*Predictive performance on the test set, CERTN*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | AUC-ROC | Sensitivity | Specificity |
| Random forests | 0.677 | 0.606 | 0.655 |
| GBM | 0.718 | 0.726 | 0.624 |
| XGBoost | 0.720 | 0.719 | 0.632 |

After weighting, per AUC-ROC, predictive performance of all three models is mediocre. The optimal random forests model—tuned to *m*try of 1—returns the lowest AUC-ROC, while the optimal XGBoost model—tuned to 1,000 rounds, maximum tree depth of 2, η of 0.01, γ of 0, sampling per tree of 0.66, and sub-sampling of 0.80—returns the highest. Essentially, each model can distinguish between certification possession and non-possession roughly seven times in every ten observations.

Regarding sensitivity and specificity, performance remains similar but mixed across models. The random forests model shows the lowest sensitivity but the highest specificity, and vice versa for the GBM model. On both metrics, the XGBoost model splits the difference between the two other models. Each model identifies true positives and true negatives in roughly six to seven out of every ten certification holders and non-holders, respectively.

The differences in predicting *CERTN* are practically negligible across models, though GBM and XGBoost slightly out-perform random forests overall. Nevertheless, selecting XGBoost based upon its AUC-ROC value, that model is applied to the validation set, which is the same size as the test set. The XGBoost model’s performance on the validation set is similar for AUC-ROC (approximately 0.710), sensitivity (approximately 0.693), and specificity (approximately 0.642).

A small set of binary features is particularly important in informing the XGBoost model. Earnings over the past 12 months—specifically, earning at least some dollar amount—proved important across all trees and boosting iterations. Also relatively important are possession of a professional license (importance score of approximately 84.427) and possession of, at most, a high school diploma (approximately 43.233). Numerous features show importance scores between roughly five and ten. Table 3 lists the top ten most important features by score, rounded to three decimal places.

**Table 3**

*Most important features in weighted XGBoost model, CERTN*

|  |  |  |
| --- | --- | --- |
| Feature | Description | Importance score |
| EEARN.L | Earning $0 to 10K over past year | 100.000 |
| LICENSE | Possession of a professional license | 84.427 |
| EDUATTN.L | Possession of, at most, a high school diploma or GED | 43.233 |
| PSCERT | Possession of a postsecondary certificate | 10.429 |
| EEWKS.L | Working, at most, 13 weeks over past year | 9.753 |
| WEPROG.Never | Never participated in a work experience program | 8.823 |
| EEEARN.C | Earning $20K to 30K over past year | 8.778 |
| EEUNION.No | Not a union member over past year | 8.366 |
| EDUFOS.Healthcare | Area of focus of highest credential is healthcare | 8.146 |
| EEWKS.C | Working 27 to 39 weeks over past year | 7.763 |

The performance of all three models, including XGBoost, does not provide strong support for using them to predict possession of a postsecondary certification. Case weighting clearly affects performance, and negatively, depending upon perspective. The non-weighted models predict that all observations, or individuals, do not possess a certification. These universal predictions are incorrect in general—roughly seven percent of individuals are certification holders—but correct for most observations. Given the ATES data are nationally representative of the adult population in the United States, addressing the imbalanced classes during training via sampling methods would not be appropriate.

By extension, those features identified as important to the selected XGBoost model may not be important in predicting actual possession of a certification. That earning at least some recent wages or salary, or earning none, could be important is no surprise given certifications, like academic or professional credentials broadly, are typically required by employers. This reality was in part the impetus for establishing the 60 by 25 goal in Illinois. However, nearly all working-age adults in the United States earn wages or salary—and strong majorities of adults have at least a credential, secondary or postsecondary—so these features may not be useful for homing in on certificates specifically. Thus, this analysis suggests that a different data set, perhaps containing individual-level data from Illinois state agencies, could better capture the differences between certification holders and non-holders.

***What characteristics predict possession of a postsecondary certificate?***

Modeling of *PSCERT* follows a similar course, beginning with a series of non-weighted models. As with postsecondary certifications, only a small proportion of the adult population in the United States possesses a postsecondary certificate. This proportion in the ATES data set is roughly nine percent.

Here again, the non-weighted models avoid distinguishing between certificate holders and non-holders. The non-weighted random forests, GBM, and XGBoost models all predict *PSCERT* in the test set with roughly 91 percent accuracy. Each optimally tuned model predicts that all observations are a “no”, or that zero individuals in the data set possess a certificate, leaving the rough nine percent of observations representing actual certificate holders as false negatives. Such predictions may be reasonable depending upon the application of such a model, but they are nonetheless not the goal of this analysis.

The second round of modeling employs a combination of case weighting during training and use of AUC-ROC during evaluation. First, case weights of roughly 9.60 applied to holders and 1.00 applied to non-holders are applied during training of each model. Second, AUC-ROC is used as the primary evaluation metric. Table 4 contains the AUC-ROC, sensitivity, and specificity metrics, rounded to three decimal places, for each weighted model.

**Table 4**

*Predictive performance on the test set, PSCERT*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | AUC-ROC | Sensitivity | Specificity |
| Random forests | 0.709 | 0.554 | 0.764 |
| GBM | 0.733 | 0.639 | 0.718 |
| XGBoost | 0.732 | 0.637 | 0.718 |

After weighting, AUC-ROC for all three models is similar. The optimal GBM model–tuned to 500 trees, maximum tree depth of 7, shrinkage rate of 0.01, and a minimum of 10 observations per node—scores approximately 0.733, while the optimal XGBoost model—tuned to 1,000 rounds, maximum tree depth of 2, η of 0.01, γ of 0, sampling per tree of 0.66, and sub-sampling of 0.80—scores approximately 0.732. The optimal random forests model— tuned to *m*try of 1—follows relatively close behind at approximately 0.709.

Sensitivity and specificity are essentially identical for the GBM (approximately 0.639 and 0.718, respectively) and XGBoost models (approximately 0.637 and 0.718, respectively). As was the case in predicting *CERTN*, here for *PSCERT*, the random forests model—tuned to *m*try of 1—returns the lowest sensitivity but the highest specificity. Also notable is specificity exceeding sensitivity for both the GBM and XGBoost models—the reverse of the case with *CERTN*. Each model identifies roughly seven out of every ten certificate holders, and roughly five or six out of every ten non-holders.

Overall, the three models again show similar but middling performance when predicting an imbalanced dependent variable—this time, *PSCERT*. Per AUC-ROC, the optimal GBM model is applied to the validation set. Its performance on that set is like the same on the test set, at approximately 0.732 for AUC-ROC, approximately 0.648 for sensitivity, and approximately 0.711 for specificity.

Compared to the features identified as important in predicting *CERTN*, the set for the GBM model have stronger scores for *PSCERT*. Most notable are working two jobs (approximately 100.000) or one job (approximately 91.419) the prior week. The rest of the top five include working 27 to 39 weeks over the past year (approximately 60.125) and using the internet for work either several times per week (approximately 46.990) or several times per year (approximately 40.102). Scores for ten most important features all exceed eleven. Table 5 lists the top ten most important features by score, rounded to three decimal places.

**Table 5**

*Most important features in weighted GBM model, PSCERT*

|  |  |  |
| --- | --- | --- |
| Feature | Description | Importance score |
| EEJOB.C | Worked two jobs last week | 100.000 |
| EEJOB.Q | Worked one job last week | 91.419 |
| EEWKS.C | Working 27 to 39 weeks over past year | 60.125 |
| XXINTFREQ.4 | Use internet for work a few times per week | 46.990 |
| XXINTFREQ.Q | Use internet for work a few times per year | 40.102 |
| EEEARN.C | Earning $20K to 30K over past year | 32.413 |
| XXINTFREQ.L | Never use internet for work | 28.986 |
| EEWKS.5 | Worked 48 to 49 weeks over past year | 17.964 |
| EEJOB.5 | Worked four jobs last week | 16.623 |
| XXMIL.NeverServed | Never served in the military | 11.967 |

None of the three models, including GBM, predict possession of a postsecondary certificate well. Though case weighting addresses the non-weighted models’ universal predictions that all observations are “no” for possession, it leaves seemingly mediocre performance. Regardless, over- or under-sampling certificate holders during model training would not reflect the true state of the population, which is well described by the nationally representative ATES data.

Interpreting the importance of binary features to the model’s predictions is challenging given the wholly categorical, sometimes ordinal nature of the variables available in the ATES data. Regarding the GBM model predicting *PSCERT*, both number of jobs worked last week and frequency of internet use post multiple features in the top ten. These general variables could be indicative of possession of a certificate or other credential, but it is unclear why, for example, working four jobs last week is more important than working three, or using the internet several times per year is more important than using it several times per month. In fact, none of the ten most important features originate as dummies; all result from one-hot encoding during data pre-processing. Further, the lack of clarity around different categories makes assessing practical significance difficult. This analysis focuses on prediction, but future work should consider focusing on explanation—as well as incorporating different data—to better understand how variables, and their categories, relate to certificate possession.

***What characteristics predict possession of a postsecondary academic degree?***

Compared to *CERTN* and *PSCERT*, PSDEGREE is balanced between classes. Degree holders make up roughly 48 percent of observations in the ATES data set, with non-holders making up the remaining 52 percent. This relative balance is to be expected given the category includes associate, bachelor’s, and graduate/professional degrees. It also renders some of the prior challenges moot and thus facilitates modeling.

An initial round of non-weighted models results in strong predictive performance on the test set. However, for consistency, case weights—roughly 1.08 applied to degree holders and 1.00 applied to non-holders—are applied in a second round, which shows slightly better performance across models. Accuracy—the proportion of predictions that are correct—and kappa—a relation between observed and expected accuracy values of the classifier—are appropriate given the balanced classes. AUC-ROC values are provided as well. Table 6 contains the three metrics, rounded to three decimal places, for each weighted model.

**Table 6**

*Predictive performance on the test set, PSDEGREE*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Kappa | AUC-ROC |
| Random forests | 0.842 | 0.684 | 0.925 |
| GBM | 0.850 | 0.700 | 0.929 |
| XGBoost | 0.854 | 0.709 | 0.934 |

Performance predicting *PSDEGREE* in the test set is similar across the three weighted models, though the optimal XGBoost model shows the highest value for each metric. Tuned to 1,000 rounds, maximum tree depth of 6, η of 0.01, γ of 0, sampling per tree of 0.66, and sub-sampling of 0.80, its accuracy is approximately 0.854, its kappa is approximately 0.709, and its AUC-ROC is approximately 0.934. The model accurately distinguishes between degree holders and non-holders for roughly 85 out of every 100 observations. Further, for a naïve comparison, its AUC-ROC value is substantially higher than the same for the optimal XGBoost model predicting *CERTN* and the optimal GBM model predicting *PSCERT*.

Delving more deeply into this better performing model, Table 7 shows the confusion matrix for the weighted XGBoost model’s validation set predictions of *PSDEGREE*. The model’s accuracy (approximately 0.850), kappa (approximately 0.702), and its AUC-ROC (approximately 0.936) all slightly lag their test set counterparts. Its sensitivity of approximately 0.918 exceeds its specificity of approximately 0.788, suggesting that the model does worse in identifying degree non-holders.

**Table 7**

*Confusion matrix for weighted XGBoost model’s validation set predictions, PSDEGREE*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Reference | |
| No | Yes |
| Prediction | No | 1,950 | 188 |
| Yes | 526 | 2,107 |

Few features informing the weighted XGBoost model’s performance show high importance scores. By far the most notable, appearing across all trees and boosting iterations, is the feature indicating no identified area of focus for the highest credential. Its high importance seems reasonable given that essentially all degree programs are either explicitly focused on an area of focus or study or require declaring a major, concentration, or the like. Next is the feature indicating an undecided area of focus, though its score of approximately 17.997 is substantially lower. Earning at least some wages or salary over the past year follows, at approximately 10.363. The remaining seven features in the top ten have scores below 10.000; Table 8 lists the top ten by score, rounded to three decimal places.

**Table 8**

*Most important features in weighted XGBoost model, PSDEGREE*

|  |  |  |
| --- | --- | --- |
| Feature | Description | Importance score |
| EDUFOS.None | No identified area of focus of highest credential | 100.000 |
| EDUFOS.GenUnd | Area of focus of highest credential is undecided | 17.997 |
| EEEARN.L | Earning $0 to 10K over past year | 10.363 |
| WEPROG.Completed | Completed a work experience program | 8.754 |
| XXINTFREQ.L | Never use internet for work | 5.180 |
| PSCERT | Possession of a postsecondary certificate | 4.563 |
| AGECAT.L | Aged 16 years to 24 years | 3.718 |
| APPRENT.No | Not completed an apprenticeship program | 2.594 |
| EDUFOS.Education | Area of focus of highest credential is education | 2.215 |
| WEPROG.Never | Never participated in a work experience program | 1.817 |

In general, despite the lower importance scores, these features make sense. For example, the traditional degree-seeking student is aged 16 to 24, though anecdotal evidence reveals adult learners are increasingly common in Illinois [and elsewhere]. Likewise, work experience and apprenticeship programs are commonly associated with career and technical education, which is not necessarily geared towards postsecondary degree-seeking. Regardless, the low scores suggest that, for this model and this data set, few features are broadly important in predicting whether an individual has a degree.

Per the solid performance of the weighted XGBoost model in predicting *PSDEGREE*, the ATES data set seems to do a better job of distinguishing possession of an academic degree versus the same for possession of either a certification (*CERTN*) or certificate (*PSCERT*). Specifically, the existence of focus area information, e.g., Education or None, is particularly key and relevant to academic degrees. However, the relative balance between degree holders and non-holders also facilitates modeling of *PSDEGREE*, whereas the imbalance of *CERTN* and *PSCERT*, respectively, poses challenges.

**Conclusion**

State of Illinois stakeholders seek to increase postsecondary credential attainment among adult Illinoisans to better meet employer demand and strengthen the state’s economy. This analysis applies classification methods to the ATES component of the publicly available, nationally representative NHES microdata to predict whether an individual possesses certain postsecondary credentials—including certifications (*CERTN*), certificates (*PSCERT*), and academic degrees (*PSDEGREE*)—and what individual characteristics inform those predictions. Such information could be key as stakeholders plan and implement the education-and-workforce-training systems that will drive credential attainment.

The predictive performance of the analysis’ classification models—random forests, GBM, and XGBoost—is mixed. Relatively few adults in Illinois or the United States possess postsecondary certifications or certificates, and this reality, captured by the ATES data, poses challenges when attempting to predict *CERTN* and *PSCERT*. Addressing the issue through case weighting results in mediocre performance that, in some contexts, may be less practical than the universal predictions of non-possession made by non-weighted models. By contrast, academic degree possession is much more balanced in the U.S. population and the ATES data set, facilitating classification and thus enabling stronger performance when predicting *PSDEGREE* relative to the other two credential dependent variables.

Across credentials, features most important in informing the relevant model’s predictions are difficult to interpret. The ATES data set used for analysis is entirely categorical, requiring conversion to dummy features prior to modeling. These dummies carry less practical value when interpreted in a vacuum. However, individual employment characteristics like earnings over the prior year are identified as important for each credential, as is frequency of internet use on the job. Regarding *PSDEGREE*, the lack of a credential area of focus—that is, an individual’s highest credential does not carry a specific focus, major, or concentration—proves particularly information, as does participation in a work experience or apprenticeship program. Notably, few if any demographic features showed great importance across models.

This analysis carries several caveats and concerns. First, it focuses on prediction over explanation. There is no consideration of how individual characteristics relate to credential possession. Predictive models could be useful for Illinois or other stakeholders, but better understanding of underlying relationships would likely be more informative. Second and related, the reliance on dummy features hinders interpretation of importance. The random forests, GBM, and XGBoost algorithms—like many machine learning algorithms—require numerical inputs, meaning the wholly categorical ATES data set needs conversion prior to modeling. The resulting dummy features, in some cases representing a single category of an ordered sequence, may be identified as important, but the practical value of that information is questionable. And third, the analysis is not Illinois-centric. A dearth of research and data focuses on the individual characteristics related to and predictive of credential attainment in Illinois. State stakeholders continue to look elsewhere for evidence, and the nationally representative ATES data set is a useful source of information. However, this data set suffers from limitations in format and timeliness that could be avoided through use of secondary, postsecondary, and workforce microdata integrated across State of Illinois data sources. Though a historical challenge in Illinois, such integration is increasingly possible via improved State infrastructure.

Future analyses could address these issues in several ways. They could employ a combination of explanation and prediction, through an expanded methodology, to add value for stakeholders. They could place greater focus on certification and certificate holders, specifically, to understand the characteristics, if any, notable that distinguish those groups. And they could prioritize analyzing Illinois-sourced data that are more relevant, user-friendly, and timely.

**References**

Garcia, L.M. & Bayer, A.E. (2005). *Variations between Latino groups in US post-secondary educational attainment*. Research in Higher Education, 46(5), pp. 511-533. Retrieved from https://link-springer-com.remote.baruch.cuny.edu/content/pdf/10.1007%2Fs11162-005-3363-5.pdf

Garner, C.L. & Raudenbush, S.W. (1991). *Neighborhood effects on education attainment: A multilevel analysis*. Sociology of Education, 64(4), pp. 251-262. Retrieved from https://www.jstor.org/stable/2112706

Illinois 60 by 25 Network (2021). *About*. Illinois 60 by 25 Network. Retrieved from https://www.60by25.org/about/

Illinois Department of Commerce and Economic Opportunity (2021). *Glossary*. Illinois WorkNet. Retrieved from https://www.illinoisworknet.com/WIOA/Resources/Pages/Glossary.aspx?

Kao, G. & Thompson, J.S. (2003). *Racial and ethnic stratification in educational achievement and attainment*. Annual Review of Sociology, 29, pp. 417-442. Retrieved from https://www-annualreviews-org.remote.baruch.cuny.edu/doi/pdf/10.1146/annurev.soc.29.010202.100019

Kuhn, M. & Johnson, K. (2016). *Applied predictive modeling*. Springer Science+Business Media.

Leventoff, J. (2018). *Measuring non-degree credential attainment, 50-state scan*. Workforce Data Quality Campaign. Retrieved from https://files.eric.ed.gov/fulltext/ED603101.pdf

The Lumina Foundation (2021). *We are tracking Illinois’ progress*. A Stronger Nation. Retrieved from https://luminafoundation.org/stronger-nation/report/2021/#/progress/state/IL

The National Center for Education Statistics (2018a). *Derived variables used in the NHES Adult Training and Education Survey (ATES) First Look Report (NCES 2018-103rev)*. Retrieved from https://nces.ed.gov/nhes/pdf/ates/2016\_ATES\_Derived\_Variables\_Note.pdf

The National Center for Education Statistics (2018b). *NHES:2016 Data File User’s Manual, 2016 Data Products, the National Household Education Surveys Program*. U.S. Department of Education. Retrieved from https://nces.ed.gov/pubs2018/2018100.pdf

The National Center for Education Statistics (2021). *The National Household Education Surveys Program*. U.S. Department of Education. Retrieved from https://nces.ed.gov/nhes/index.asp

Rumbaut, R.G. (2005). *Turning points in the transition to adulthood: Determinants of educational attainment, incarceration, and early childbearing among children of immigrants*. Ethnic and Racial Studies, 28(6), pp. 1041-1086. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01419870500224349?scroll=top&needAccess=true

Srivastava, T. (2016). *How to XGBoost algorithm in R in easy steps*. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/#h2\_6

State of Illinois (2018). *Illinois Career Pathways Dictionary*. Illinois State Board of Education. Retrieved from https://www.isbe.net/Documents/IL-Career-Pathways-Dictionary.PDF

Thompson, M.S., Gorin, J.S., Obeidat, K. & Chen, Y. (2006). *Understanding differences in postsecondary educational attainment: A comparison of predictive measures for Black and White students*. The Journal of Negro Education, 75(3), pp. 546-562. Retrieved from https://www.jstor.org/stable/40026821

Wilson, K.R. & Allen, W.R. (1987). *Explaining the educational attainment of young Black adults: Critical familial and extra-familial influences*. The Journal of Negro Education, 56(1), pp. 64-76. Retrieved from https://www.jstor.org/stable/229538

Wolfle, L.M. (1985). *Postsecondary educational attainment among Whites and Blacks*. American Educational Research Journal Winter 1985, 22(4), pp. 501-525. Retrieved from https://journals-sagepub-com.remote.baruch.cuny.edu/doi/pdf/10.3102/00028312022004501

xgboost developers (2021). *XGBoost R tutorial.* XGBoost R Package. Retrieved from https://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html