**Evaluating Factors Enabling Graduation with the High School Longitudinal Study of 2009**

Janette Avelar

Department of Education Studies, University of Oregon

EDLD 654: Machine Learning

Dr. Cengiz Zopluoglu

December 8, 2023

**Introduction**

The field of Machine Learning opens the possibility for large-scale data analysis that can have useful implications for the field of education research. One of its most promising strengths is that it allows for the prediction of outcomes from a much larger number of predictors than traditional statistical approaches may be able to handle. Resulting findings can help illuminate new areas for research that may otherwise be overlooked and can also help with variable selection when building models from exceptionally large datasets. In this exploratory project, I demonstrate this by using machine learning to predict graduation outcomes for high school students using a subset of the publicly available High School Longitudinal Study of 2009 (HSLS) dataset provided by the National Center for Education Statistics (NCES).

NCES creates and hosts publicly available datasets that education researchers can explore and analyze. NCES datasets tend to be very large and often track data over time, adding complexity to potential analysis and making it more difficult for researchers to identify models when they have a plethora of variables to choose from. Further, for a burgeoning data analyst, NCES datasets allow for open data exploration that can bring research questions forward.

The HSLS dataset catalogues potential indicators that may enable or hinder high school graduation and postsecondary outcomes. It is comprised of the results of student, parent, teacher, and administrator questionnaires, student transcripts, and composite variables created by NCES. These data catalogue student experiences, student characteristics, teacher and administrator credentials, and student achievement for more than 23,000 students across 10 states beginning in 2009. The study currently has three waves: the first in 2009, with an initial follow-up in 2012, and a second follow up in 2016. HSLS is a useful dataset for education researchers not only because of its breadth, but also because of its longitudinal nature. In follow-up studies, the initial high school cohort is followed through their postsecondary years and into their early 30s. This opens the possibility to explore the impact of public schooling on lifelong trajectories.

**Research Problem**

The education sector in the United States is constrained by bureaucratic and financial factors that challenge aims to successfully graduate as many students as possible. Limited resources require that we funnel money and effort into interventions and practices that hold great promise. This is especially important to balance in the current moment, as we face significant learning loss following COVID-19 and a looming fiscal cliff with the conclusion of Elementary and Secondary School Emergency Relief (ESSER) grants. In the wake of ESSER, schools are being criticized for the ways they incorporated funds to support students through the COVID-19 crisis. Increased vigilance and skepticism are prompting the need to accurately identify areas to invest in with evidence of their promise. But how should we identify areas to investigate in the first place?

One approach is to start from the ground up, identifying areas with promise we observe through practice. However, this approach may be biased, as the breadth of practices we’re likely to be exposed to in a given context is quite narrow. Another approach is to leverage existing datasets that can connect student outcomes with a much wider breadth of practices and potential confounding factors. This approach can serve as a starting point, to determine routes for investigation with some level of initial evidence. I will test out this approach in the following exploratory study by comparing two types of predictive models and a regression model to identify which factors in the HSLS student dataset most predicted students would graduate from high school. The results are intended to identify a starting point from which further studies of potential policy solutions and interventions could be designed.

**Methods**

In this study, I use a classification model with a ridge penalty to identify which factors most predicted students would graduate from high school. I then compare the performance of the classification model with a ridge penalty to a classification decision tree model, and a traditional logistic regression model.

**Description of the data**

The HSLS dataset is a nationally representative and longitudinal dataset tracking a cohort of students starting in 2009. In the first wave, students were attending 9th grade and in the most recent wave in 2016 the cohort is around 30 years old. The surveys used in the study were answered by students, parents, teachers, counselors, and administrators. These surveys are intended to better understand characteristics, for example, levels of education, types of certifications, occupation, and general attitudes or beliefs about school and college goals. Additionally, transcripts and school characteristic data can reveal how course taking and schoolwide practices may hinder or enable graduation rates.

For this study, I used the public-facing version of the most recent HSLS data from 2017. I included NCES’ composite student variables except for associated weights because I did not include weights in my preliminary models, but I expect to incorporate them in the next iteration. I excluded original survey responses because I wanted to build a starting point from more direct characteristics, though a promising next step is to include survey responses to identify attitudes and behaviors that may also promote student outcomes. I also excluded school variables because the public-facing version does not include unique identifiers to merge the school and student datasets. The final dataset was comprised of the following subsets:

* IDs and weights (with weight variables removed)
* BY student level composites
* F1 student level composites
* HS Transcript student level composites

**Data Preparation**

To prepare the dataset, I dropped all suppressed variables from the original NCES download. I then recoded all potential missing indicators across the dataset including:

* “Missing”
* Values indicating differ types of missingness (-8, -6, -9)
* “Unit non-response”
* “Item legitimate skip/NA”
* “Component not applicable”
* “Nonrespondent”

Identifying the type of missing data was not relevant for an initial exploration. I then recoded all binary variables to a standardized 0/1 format and dropped key demographic indicators for time 2 and 3 (gender, race/ethnicity, first language) that were not time varying. I also modified the outcome variable (graduation) to a binary format (more information below) and removed all observations where the outcome variable was missing. Finally, I used the **caret** package in R to build a recipe for processing data that applied one-hot encoding to all categorical variables, created missing indicators for all predictors, removed variables with zero or near-zero variance, imputed missing numeric values with the variable mean, and imputed missing values in categorical variables with the mode. Once my data was processed, I split the data into training and testing datasets using an 80/20 split.

***Outcome variable***

The outcome variable for the classification models is a binary version of the HSLS graduation outcome variable from the most recent wave of data. The original variable (X3TOUTCOME) indicates graduation outcome as provided by the high school, and specifies one of the following labels:

* Fall 2012-summer 2013 graduate\*
* Post-summer 2013 graduate\*
* Pre-fall 2012 graduate\*
* Graduation date unknown
* Certificate of attendance
* Dropped out
* Transferred
* Left other reason
* Still enrolled
* Status cannot be determined
* Unit non-response

The labels I coded as `1` indicating that a student successfully graduated, are marked with an asterisk. The label “Unit non-response” was coded as missing, and all other labels were coded as `0`.

**Figure 1**

*Rates of Graduation for HSLS Cohort in 2016 Follow-up*

*A graph showing a bar graph

Description automatically generated*

***Sample***

The original sample for the HSLS study was preserved to the extent possible. The only observations removed were those missing a value for the outcome variable. The resulting sample was predominantly White (52%), with very few American Indian and Alaska Native students (1%) and no Native Hawaiian and Pacific Islander students. It’s likely these percentages are further suppressed due to NCES efforts to ensure students remain deidentified in contexts where they are in the extreme minority.

**Figure 2**

*Racial and Ethnic Demographics of HSLS Sample*

*A graph with a bar graph

Description automatically generated*

The sample had a balanced male/female population but included no non-binary or otherwise genderqueer students. This may be due to survey design or may again result from suppressing small numbers to deidentify student data.

**Figure 2**

*Gender Demographics of the HSLS Sample*

A graph showing a number of students

Description automatically generated

**Description of the models**

In the section below, I detail the specific settings used during model fitting and the criteria I used to evaluate model performance. I used the R package **caret** to train all three models. All models were trained using the training dataset and evaluated on their performance with the testing dataset. When building a confusion matrix for each model, I set a threshold of 0.5. Ideally, in the future this threshold would be modified to maximize sensitivity in order to most accurately identify true positives, but for an initial exploration I hoped to compare similar metrics and wanted to identify a baseline performance.

***Logistic regression model***

It is not necessary to specify settings or undergo hyperparameter tuning for the logistic regression model.

***Ridge regression model***

For the classification model with a ridge penalty, I created a tuning grid that set alpha to 0, and tested values of lambda between 0 and 0.5 using increments of .001. These values were identified after a series of trial and error with much larger numbers (a range of 0 to 5, a range of 0 to 3, and a range of 0 to 1). The best fitting model had a lambda value of 0.04.

***Decision tree model***

For the classification decision tree model, I created a tuning grid that tested values of the complexity parameter between 0 and 0.5 using increments of .01. These values were identified after a series of trial and error with larger ranges. The best fitting model had a complexity parameter of 0.01. I also specified a minimum split value of 5, a minimum bucket value of 2, and a maximum depth of 10. Again, these values were identified after a series of trial and error with larger numbers. These values are also constrained by the performance of my computer, as the time needed to run models was a factor in choosing smaller numbers.

***Evaluation plan***

To determine the best performing classification model, I will compare values for the Area under the Receiver Operating Curve (AUC) for each model, along with sensitivity (TPR) and specificity (TNR) metrics. Because I am most interested in predictive power for this initial step of exploration, I will value AUC over other metrics. However, I will also use sensitivity and specificity metrics alongside overall accuracy (ACC) if AUC values are very similar. Further, I will also compare the overall R-square of each model to evaluate performance.

**Results**

Overall, I found that the classification model with the ridge penalty performed the best of the three models. The decision tree classification model had relatively similar performance, but ultimately performed just slightly worse on AUC and overall accuracy. Additionally, the model identified the ten most important predictors for graduation, which are listed in the table and figures below.

**Table 1**

*Ridge Classification Model - Ten Most Important Predictors for Graduation*

|  |  |
| --- | --- |
| **Variable Name** | *Description* |
| X3TSTATYR13 | Indicator for extended schooling at time point 3 representing if a student took at least one course in the 2013/14 school year. Most students in the cohort graduated at the end of the 2012/13 school year, suggesting extended time to graduation. |
| X3TGPAMTHAP\* | Student GPA in AP or IB Mathematics courses for 2015/16. The modified version of the variable indicates whether the value was missing. |
| X3MOMOCC2\* | For participants who have a biological, adoptive or step mother living in their household, the 2-digit Occupational Information Network (O\*NET) code for the most current job identified at time point 2. The modified variable indicates mother or female guardian’s most recent occupation is in the military. |
| X3TGPAPUBSER\* | Student GPA in the course Public, Protective, and Government Service. The modified variable indicates whether the value was missing. |
| X1DADOCC\_STEM1\* | For participants who have a biological, adoptive, or step father living in their household with a STEM occupation, the sub-domain within which their most recent occupation at time point 1 was held. The modified variable indicates father or male guardian’s most recent STEM occupation was in an unspecified sub-domain. |
| X3TAGPA12\* | GPA in twelfth grade academic courses at time point 3. The modified variable indicates whether the value was missing. |
| X3TTRNLASTHS | Indicator that a transcript was provided for the last known school as the transcript component at time point 3. |
| X1TSRACE\* | Characterizes the race/ethnicity of the participant’s science teacher. The modified variable indicates that science teacher’s ethnicity was Hispanic/Latinx with no race specified. |
| X3TGPASCIAP\* | GPA in AP or IB Science courses for 2015/16. The modified version of the variable indicates whether the value was missing. |
| X3TGPAENG\* | GPA in English Language and Literature for 2015/16. The modified version of the variable indicates whether the value was missing. |

\*indicates a modified version of the variable developed in data processing phase

**Figure 3**

*Most Important Predictors for the Classification Model with a Ridge Penalty*

A graph with black dots and white text

Description automatically generated

**Model performance and fit**

Of the three models, the traditional logistic regression model performed the worst, with an overall accuracy of 0.82 and an AUC of 0.82. It also had the lowest true positive rate at 0.93, though this is still high, and the highest true negative rate of 0.7, indicating it was best able to accurately predict when a student did not graduate as compared to the other models (ridge = 0.66; tree = 0.61). The classification tree model performed similarly in terms of overall accuracy (0.8) and AUC (0.82). It held the highest true positive rate of all three models, indicating it was best able to predict when a student graduated (0.99). However, as I was most interested in predictive power, I ultimately decided the classification model with a ridge regression outperformed the tree model, with an AUC value of 0.91, despite a slightly worse true positive rate of 0.98. The model fit metrics are summarized in the table below.

**Table 2**

*Classification Model Performance Summary*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **ACC** | **TPR** | **TNR** | **PRE** |
| *Logistic Regression* | 0.82 | 0.82 | 0.93 | 0.71 | 0.76 |
| *Ridge Classification* | 0.91 | 0.82 | 0.98 | 0.67 | 0.75 |
| *Classification Tree* | 0.82 | 0.80 | 0.99 | 0.61 | 0.72 |

**Discussion**

After identifying the classification model with a ridge penalty as the model of choice, I used the **vip** package in R to extract the most important predictors in the model (see Figure 3 below). Half of the predictors identified were missing flag indicators generated when processing the data. This includes predictors that indicate missing GPA scores for English Language and Literature, AP or IB Science, AP or IB Mathematics, overall twelfth grade academic courses, and what appears to be a Civics course titled Public, Protective and Government Service. The fact that these missing indicators attributed so greatly to predicting graduation could indicate that involvement in certain core courses (Language Arts, Science, Math, Social Studies) is critical to ensuring timely graduation. This is not surprising given that high schools greatly rely on core content course enrollment and performance to determine graduation. It is also interesting to note that the level of coursework played a role, which may suggest that enrollment in advanced coursework like AP or IB helped to support timely graduation. It is similarly not surprising that missing a GPA score for the entire twelfth grade year would be an important predictor, as it might indicate a student who dropped out or was experiencing circumstances that prevented them from attending school in twelfth grade. The indicator X3TTRNLASTHS which specifies whether a transcript was provided for the last known high school may similarly suggest a missing or incomplete transcript that a school was not able to hand over, or it may indicate a school that was less likely to participate. An important next step before relying on these conclusions is to investigate whether a missing value was likely to indicate a student was not enrolled in the specified coursework and whether a missing transcript is reflective of whether a student was unable to complete coursework at the school, or if the school failed to hand over transcripts for certain participants for other reasons.

On the other hand, while the variables discussed above suggest that what a student does not have access to may greatly impact graduation, the variable X3TSTATYR13 suggests that what a student *does* access may also play a role. This indicator specified whether a student was enrolled in at least one course following the year that a typical student in the cohort graduated. This means that students who extended the length of their schooling were also more likely to graduate, and it would be interesting to explore how this variable correlates with other factors that may play a role in extended learning time. For example, students who are classified as English learners and students eligible for Special Education are two populations that commonly receive extra schooling through their secondary years. If the majority of the students predicted to graduate fall into these groups, it may provide evidence that extended schooling time is supporting their ability to graduate and also that some work could be done to ensure these students are able to graduate on time with their peers to minimize the need for extended schooling. This work is especially important given this variable was one of the top two predictors with much greater importance than some of the other predictors in the top ten.

Finally, the remaining predictors were related to parents’ occupation and the ethnicity of the student’s science teacher, illuminating areas of further study. First, whether the mother or female guardian’s most recent occupation was in the military had an impact on predicting graduation. This may be due not necessarily to the occupation itself, but to the services and schooling provided as a result. It may be the case that these students were enrolled in schools operated by the U.S. military which may operate significantly differently than the typical U.S. public school. Additionally, U.S. military benefits may ensure that these students are part of a higher socioeconomic class and have access to more stable environments. Further study on the sample may illuminate differences that are adaptable to the larger U.S. public education sector.

Secondly, whether the father or male guardian’s most recent occupation was in an unspecified sub-domain of STEM also attributed to predicting graduation. In addition to further analyzing the sample to understand why this is the case, it may also be interesting to investigate whether occupation correlated with other STEM-related outcomes, for example, if students were more likely to enroll in certain courses or to pursue a STEM degree or career. Finally, the ethnicity of the student’s science teacher, and specifically whether they identified as Latinx, also impacted predicting graduation. Again, it may be worth further investigating which students this most impacted to draw inferences and potential policy recommendations. Though speculative, it may be the case that dimensions of racial matching, where students perform better in school when the instructional staff they interact with is from the same racial or ethnic background as them, are at play or could potentially support theories that a racially diverse teacher population contributes to higher student achievement and graduation outcomes, particularly for students of color.

**Conclusion**

I used machine learning to compare the predictive power of classification models against a traditional logistic regression model. The machine learning algorithms outperformed the traditional logistic regression model, and the classification model with a ridge penalty held the most predictive power. The extracted top predictors from the ridge model suggested the importance of coursework, or rather, the lack of certain coursework on a student’s transcript, as a critical indicator for graduation. Additionally, the model revealed new potential areas of study to investigate potential practices or policies that may support high school graduation, including extended learning time, the racial diversity of instructional staff (and in particular, science teachers), and increasing access to core content and advanced coursework. This exploratory paper details the process for how a machine learning algorithm can be used to identify potential research questions, areas of study, and variables for further investigation from a large dataset that may otherwise be difficult to tackle and specify models for. This may be an especially useful process for graduate students learning to use data science to inform education research and are thus working with large-scale datasets.