David Braslow Capstone Project

Machine Learning Engineer Nanodegree

I. Definition

Project Overview

Science, technology, engineering and mathematics (STEM) education has received renewed interest in the USA as people and organizations have become increasingly reliant on computers and other advanced technologies. Labor market demand for workers with technical skills often outstrips supply, and wages for STEM jobs tend to be high and are expected to rise. While this could be an opportunity for many disadvantaged students to find high-paying jobs and improve their life prospects, they often experience low performance in or don't have access to secondary STEM courses. As a result, disadvantaged students have low representation in STEM post-secondary programs and make up only a small percentage of STEM graduates.

Despite the challenges they face, some disadvantaged students do go on to pursue post-secondary STEM education. We can learn from these students what factors are most important for their ongoing interest in STEM. Increasing the STEM attainment of disadvantaged students in post-secondary programs is an important goal not only for the industries that require a robust supply of STEM college graduates, but also for efforts to improve the quality of life for disadvantaged students.

For this study, I use data from the High School Longitudinal Study, 2009-2013 (HSLS:09) conducted by NCES[1]. The study follows a nationally representative group of high school students through high school, recording a number of student, school, and family variables. In total, 23,503 students responded from 944 high schools.

The sample of interest – low-income students – are defined as those from families with household income below 185% of the Census poverty threshold (5,558 students). The inputs of interest will be credits earned in various specific STEM courses, total credit earnings in various STEM disciplines (e.g. math, biology, engineering), and GPA in

[1] United States Department of Education. Institute of Education Sciences. National Center for Education Statistics. High School Longitudinal Study, 2009-2013 [United States]. ICPSR36423-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-05-12. http://doi.org/10.3886/ICPSR36423.v1

various STEM disciplines. The outcome of interest will be whether the student is enrolled in a postsecondary program as of November 1, 2013 and considering a STEM major, which I call *Postsecondary STEM Pursuit* (PSTEMP).

Problem Statement

The problem I aim to solve is to determine which high school STEM experiences are most important for predicting whether low-income students want to pursue STEM post-secondary education. To answer this problem, I will train a decision tree to classify students by postsecondary STEM pursuit using the inputs described. I will use the approach developed by Garson (1991)[2] to identify the most important variables in the neural network.

Metrics

The evaluation metric I use is F1 Score. This is appropriate because the outcome of interest is skewed and because it incorporates both recall and precision. I have no reason to weight one over the other, so a balanced F score is used.

[2] Garson, G.D. 1991. Interpreting neural network connection weights. Artificial Intelligence Expert. 6(4):46–51.

II. Analysis

Data Exploration

One unusual feature of this dataset is that different categories of missingness are coded into each variable. For example, the question asking students whether they intend to pursue a STEM major has separate codes for "Don't Know", "Item not administered: abbreviated interview", "Item legitimate skip/NA", "Unit non-response", and "Missing". These missing values comprise most (51%) of the values for this variable.

Among the 4,020 remaining students, there is a wide range in STEM experiences and outcomes. With regards to course taking, HSLS asks about coursework in 8th, 9th, and 12th grade. It asks about the most advanced math and science courses taken in 8th grade:

```
    S1 B06 Most advanced math |
course taken by 9th grader in |
             the 8th grade |      Freq.
---------------------------+-----------+-----------------------------------
                   Math 8 |        928 |*********************
Advanced or Honors Math 8 |         91 |**
              Pre-algebra |      1,520 |***********************************
Algebra I including IA and IB |        919 |*********************
Algebra II or Trigonometry |         27 |*
                 Geometry |         91 |**
           Integrated Math |         87 |**
         Other math course |        216 |*****
---------------------------+-----------+-----------------------------------
                    Total |      3,879
```

S1 B08 Most advanced science course taken by student in the 8th grade	Freq.	
Biology	82	*
Life science	381	*******
Pre-AP or pre-IB Biology	18	
Chemistry	36	*
Earth Science	584	**********
Environmental Science	79	*
Integrated Science	45	*
General Science or General Science 8	318	******
Science 8	1,577	*****************************
Physical Science	464	********
Physics	29	*
Other science course	212	****
Total	3,825	

HSLS asks whether students are enrolled in the following courses in fall of 9th grade:

Math (N = 3,432)		**Science (N = 3,072)**	
Algebra I (including IA and IB)	61.0%	Biology I	35.4%
Geometry	17.4%	Earth Science	15.8%
Algebra II	5.8%	Physical Science	26.9%
Trigonometry	0.2%	Environmental Science	4.7%
Review or Remedial Math	1.2%	Physics I	3.5%
Integrated Math I	4.2%	Integrated Science I	4.8%
Statistics or Probability	0.3%	Chemistry I	2.9%
Integrated Math II or above	0.6%	Integrated Science II or above	0.3%
Pre-algebra	8.0%	Advanced Biology	1.7%
Analytic Geometry	0.1%	General Science	2.6%
Other advanced math course	0.3%	Life Science	2.4%
Other math course	7.2%	Advanced Physics	0.3%
		Other earth/environmental science	0.4%
		Other biological science	0.2%
		Other physical science	0.2%
		Other science	7.4%

HSLS asks whether students are enrolled in the following courses in spring of 12th grade:

Math (N = 2,950)		**Science (N = 2,675)**	
Pre-Algebra	2.6%	Life Science	1.6%
Algebra I (Including IA And IB)	8.6%	Biology I	12.1%
Algebra II	40.9%	Biology II	4.3%
Algebra III	4.9%	Advanced Placement (AP) Biology	2.4%

Geometry	20.1%	International Baccalaureate (Ib) Biology	0.4%
Analytic Geometry	0.6%	Anatomy Or Physiology	6.0%
Trigonometry	9.2%	Other Biological Science Courses	5.6%
Pre-Calculus Or Analysis And Functions	16.7%	Chemistry I	36.9%
Advanced Placement (AP) Calculus AB Or BC	2.1%	Chemistry II	4.1%
Calculus Other Than AP	1.0%	Advanced Placement (AP) Chemistry	1.9%
Advanced Placement (AP) Statistics	1.0%	International Baccalaureate (IB) Chemistry	0.2%
Statistics Or Probability Other Than AP	3.4%	Earth Science	6.6%
Integrated Math I	1.6%	Advanced Placement (AP) Environmental Science	1.8%
Integrated Math II	1.4%	Other Earth Or Environmental Science	4.2%
Integrated Math III Or Above	2.1%	Physics I	16.0%
Business/General/Applied/Technical/Review Math In	4.8%	Physics Ii	1.5%
Other Math Course	8.7%	Advanced Placement (AP) Physics B Or C	1.5%
		International Baccalaureate (IB) Physics	0.2%
		Physical Science	6.2%
		Other Physical Science	0.8%
		Integrated Science I	0.8%
		Integrated Science II Or Above	0.4%
		General Science	1.1%
		Computer Applications	3.4%
		Computer Programming	1.4%
		AP Computer Science	0.2%
		Other Computer Or Information Science Course	1.9%
		Engineering	2.1%

Additional course-taking variables include the following:

Highest level mathematics course taken/pipeline
Highest level mathematics course taken - ninth grade
When student took Algebra I
Highest level science course taken
Highest level science course taken - ninth grade
Highest level biology course taken/pipeline
Highest level chemistry course taken/pipeline

Highest level physics course taken/pipeline
Highest level other science course taken/pipeline
Has taken an AP math course(s)
Has taken an AP science course(s)
Has taken IB math course(s)
Has taken IB science course(s)
Has taken math dual enrollment course(s)
Has taken science dual enrollment course(s)

HSLS asks about whether students earned at least one credit in the following STEM subjects by the spring of 12th grade (N = 3,818):

Algebra 1	91.6%
Algebra 2	52.5%
Integrated Math	8.2%
Analysis/Pre-Calculus	20.0%
Calculus	9.3%
Geometry	69.5%
Statistics/Probability	6.4%
Trigonometry	10.6%
Biology	82.5%
Chemistry	53.0%
Geology/Earth Science	69.1%
Physics	25.7%

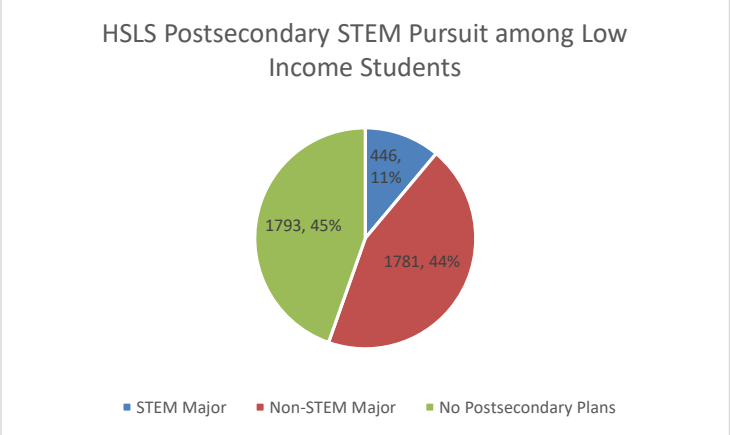
HSLS also asks about the number of credits earned in various courses:

Credits earned in: AP/IB mathematics courses
Credits earned in: mathematics
Credits earned in: AP/IB science courses
Credits earned in: science

In total, there are 132 features in this dataset. A sample of the dataset, including four features and the target, is presented below:

S2ENGINEER12	S2HISCIENCE12	S3CLASSES	S3CLGFT	S3FIELD_STEM
0	9	1	1	0
0	6	1	1	0
0	8			0
0	15	0		0
0	17	0		0
0	3	1	1	0
0	8	0		0
0	8	1	1	0
		1	1	1
		1	1	0
0	8			0
		0		0
		0		0
		0		0
		0		0
0	16	1	1	0
0	8	1	1	0

Exploratory Visualization



This visualization shows the various postsecondary STEM outcomes for low-income students in the HSLS dataset. It shows that of the 4,020 low-income students studied, about 45% do not pursue postsecondary education, 44% pursue non-STEM postsecondary education, and 11% pursue STEM postsecondary education. This shows that it is rare for low-income students to pursue post-secondary STEM majors, even after accounting for those who do not pursue postsecondary education of any kind.

Algorithms and Techniques

To answer the problem as stated above, I will train a decision tree to classify students by postsecondary STEM pursuit using the inputs described. I chose a decision tree because it is appropriate for supervised learning with dichotomous outcomes, when there are a large number of feature variables, and when some variable are non-binary. The target will be the post-secondary STEM pursuit variable, operationalized as a dichotomous outcome (PSTEMP = 1, all other outcomes/No PSTEMP = 0).

The decision tree algorithm I will use is the default algorithm for decision trees in scikit-learn, which is the CART algorithm. It begins by identifying the feature and threshold that provide the greatest information gain, and uses them to create a rule that bifurcates the sample. Then for each subsample, the algorithm repeats this step. This continues until there is no further information gain is found or the maximum number of allowed steps have been completed.

Benchmark

Among the 4,020 low-income students whose Postsecondary STEM Pursuit we know, only 446 (11.1%) are pursuing post-secondary STEM education. I will compare the testing F1 score from my model when predicting the "No PSTEMP" class to the testing F1 score of a naïve classifier that randomly assigns 88.9% of students to the "No PSTEMP" class.

III. Methodology

Data Preprocessing

For the purpose of this study, I include students as "No" observations for Postsecondary STEM Pursuit if they were not asked the relevant question because they were not enrolled in post-secondary classes (which corresponds to the "Item legitimate skip/NA" option). All other values were coded as missing, and the 1,538 students with missing values on this variable (27%) were dropped from the dataset due to lack of observed target.

Other feature variables were also coded as missing using similar logic, but no further students were dropped from the dataset. Missing values were imputed as either the mean or as zero, depending on the reasons for missingness.

Some of the variables in this dataset are dichotomous, but many are not. The credit variables are count variables, the GPA variables are continuous, and the "highest level course" variables are ordinal. The decision tree algorithm can handle these variables by setting threshold values. The "most challenging course" variable, however, is not ordinal but categorical, and was thus converted into dummy variables before analysis.

Implementation

Some pre-processing steps were implemented in Stata 14, while the rest of the methods described were implemented in Python 3.6 using numpy (v. 1.11.1), pandas (v. 0.18.1) graphviz (v. 0.5.2) and scikit-learn (v. 0.17.1). The data were split into a training set with 3,000 observations and a test set with 1,020 observations. Two classifiers – a decision tree with default settings and a naïve classifier with an assignment probability of 88.9% - were trained on the training set. The default settings for a decision tree in scikit-learn use the gini coefficient to make decision rules, set no maximum numbers for depth, nodes, or variables, and set the minimum samples for splitting at 2. The classifiers were then applied to the test set, and the F1 score was calculated and compared.

Refinement

My initial implementation had a testing F1 score of 0.9068. I refined my decision tree classifier by testing different maximum depths and different values for the minimum number of samples required to create a split in the tree. The original implementation did not set a maximum depth and set the minimum number of samples for splitting at 2.

However, since there are many features in my training set, changing these parameters may improve performance by preventing the decision tree from using features that have low incidence or importance in the full population, but that happen to be moderately important in the training data. I tested maximum depths of 3 to 10 and minimum sample numbers from 3 to 30. The resulting implementation had a testing F1 score of 0.9387.

IV. Results

Model Evaluation and Validation

The final model was chosen based on the highest F1 score. I refined this model by testing different maximum depths. To get the highest training F1 score, the maximum depth was 4, and the minimum sample for splitting was 26. I further tested the sensitivity of this model by using k-fold cross validation, using 10 folds. The range of F1 scores from this analysis was somewhat wide (min = 0.9314, max = 0.9564, mean = 0.9434), suggesting that the model is not very sensitive to the chosen training set, and the mean is close enough to our F1 value and to 1.0, suggesting that the model is trustworthy.

Justification

The F1 score of the tuned model on the test set was 0.9387, which is greater than the F1 score for the benchmark model on the test set (0.8896). This value suggests there is minimal misclassification. This is unsurprising because the low incidence of post-secondary STEM pursuit is makes it easy to avoid misclassification by erring on the side of assigning "No PSTEMP" labels.

Further, the tree that was generated relies on variables that, on their face, seem likely to be important for students' post-secondary decision making. The first split – whether students are taking post-secondary classes at all – is a deterministic predictor in this dataset, since students who were not taking classes were not asked about their interest in STEM majors. The second split – whether a student took Calculus in high school – also makes sense, since Calculus is typically optional for high schoolers, so taking it indicates interest in and ability to succeed in math.

V. Conclusion

Free-Form Visualization

The figure below (Figure 1) shows the tuned decision tree from this analysis. Each box contains five pieces of information (in order from top to bottom):

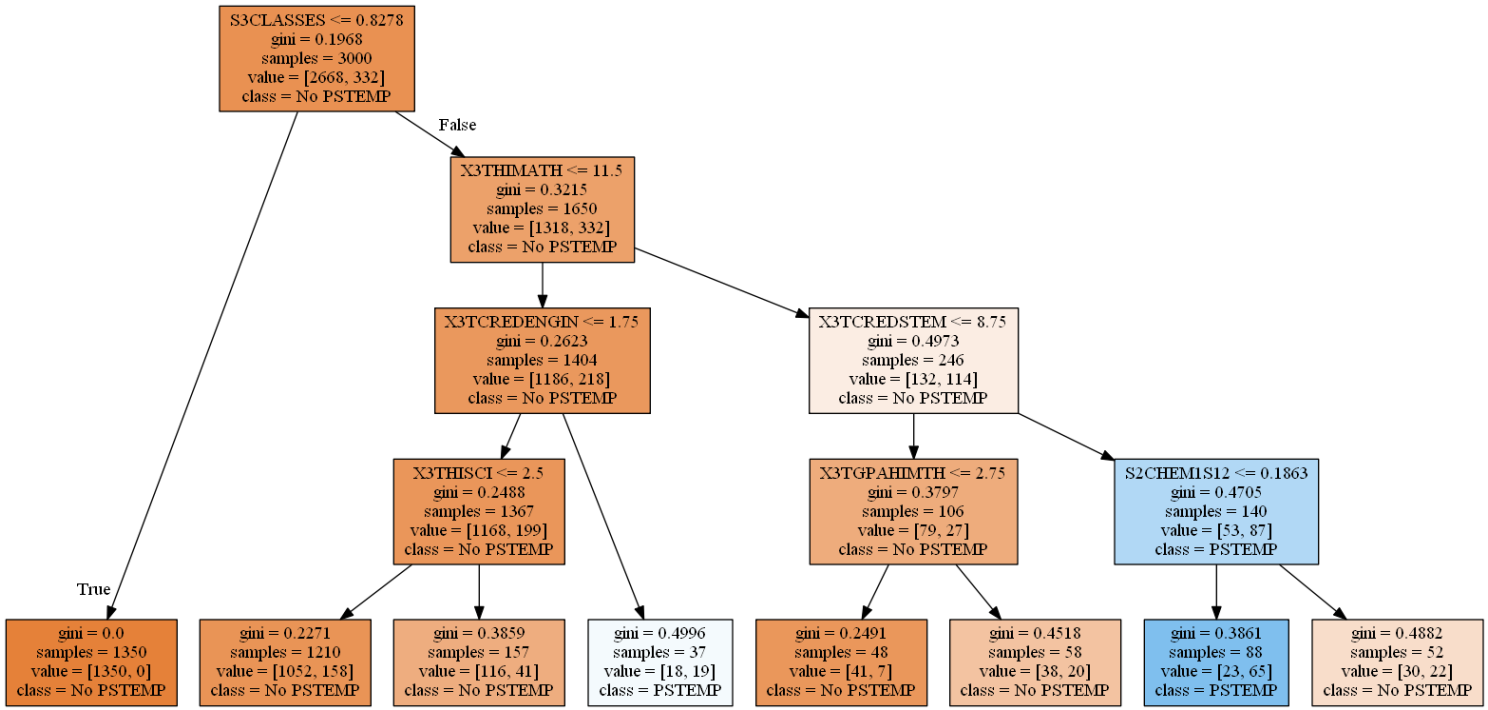
1. The rule used to decide which path to follow, including the variable and the boundary value.
2. The gini coefficient of the sample at that node from the training set.
3. The number of students in the sample at that node from the training set.
4. The number of students with each observed value of the target at that node from the training set (left = No PSTEMP, right = PSTEMP).
5. The modal observed target of the students at that node from the training set.

Key:

- S3CLASSES: Taking postsecondary classes as of Nov 2013 (1 = Yes, 0 = No)
- X3THIMATH: Highest level math course taken (13 = AP Calculus, 12 = Calculus, 11 and under = Precalculus and below)
- X3TCREDENGIN: Number of credits earned in engineering/engineering tech
- X3TCREDSTEM: Number of credits earned in STEM
- X3THISCI: Highest level science course taken (4 = AP/IB, 3 = Advanced, 2 and under = Specialty and below)
- X3TGPAHIMTH: GPA in highest level math course taken (0-4 scale)
- S2CHEM1S12: Taking chemistry I spring 2012 (1 = Yes, 0 = No)

Commented [DB1]: Nice job here! The reason that this isn't marked as passing is that there should be some labels to indicate what the decision tree means. For instance, the reader isn't going to know what S3Classes or X3TWHENALG1 are. Please add some labels so that the reader can interpret how the algorithm is making decisions. Additionally, you should provide some discussion about what the reader should take away from this figure. Does it indicate what factors are most important, or suggest a way to help students pursue STEM education?

Figure 1: Post-Secondary STEM Pursuit Tuned Decision Tree



Reflection

The project started with the goal of predicting post-secondary STEM pursuit from high school STEM experiences using the HSLS dataset. However, the initial data exploration and variable selection turned out to be the most challenging part. From the thousands of variables in the dataset, I had to select those that I thought could be considered indicative of "high school STEM experiences". These variables also had many different scales and patterns of missingness, which required substantial pre-processing. Fitting and refining the decision tree turned out to be comparatively straightforward. Coding turned out to be a challenge as well, as I am a Python novice.

The final model fit my expectations in terms of its structure and in terms of it's F1 score. The amount of improvement from the baseline model, while small, was enough to be worth keeping the model. Therefore, I would be comfortable recommending the model's use in other situations.

Improvement

It is quite likely that a better solution exists for this problem. Given the large numbers of variables in this dataset, it is possible that other methods of pre-processing and selecting variables would lead to a better solution. Other stronger learners, such as a neural network, may have been able to better use the information from all of the feature variables to improve our F1 score. However, the simplicity of the decision tree makes it more useful for informing school personnel about high school STEM experiences to pay attention to when trying to identify low-income students who may be interested in STEM.