**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]](#footnote-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 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*.

### 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 neural network to classify students by postsecondary STEM pursuit using the inputs described. I will use the approach developed by Garson (1991)[[2]](#footnote-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.

## 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 |

### 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 and when there are a large number of feature variables. The target will be the post-secondary STEM pursuit variable, operationalized as a dichotomous outcome (STEM pursuit = 1, all other outcomes = 0)

### 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 my model to a random assignment model with an 11.1% probability of assignment.

## 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 not dichotomous. The credit variables are count variables, the GPA variables are continuous, and the “highest level course” variables are ordinal. I trichotomize these variables into roughly equal size groups to make them easier to include as features in the neural network. The “most challenging course” variable, however, is not ordinal but categorical, and was thus converted into dummy variables before analysis.

### Implementation

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 and a dummy classifier - were trained on the training set. The classifiers were then applied to the test set, and the F1 score was calculated and compared.

### Refinement

I refine my decision tree classifier by testing different maximum depths. The original implementation did not set a maximum depth. However, since there are many features in my training set, this 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 depths of 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, and 100. The depth with the highest training F1 score ended up being 15.

## IV. Results

### Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

* Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?
* Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?
* Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?
* Can results found from the model be trusted?

### Justification

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

* Are the final results found stronger than the benchmark result reported earlier?
* Have you thoroughly analyzed and discussed the final solution?
* Is the final solution significant enough to have solved the problem?

## V. Conclusion

(approx. 1-2 pages)

### Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

* Have you visualized a relevant or important quality about the problem, dataset, input data, or results?
* Is the visualization thoroughly analyzed and discussed?
* If a plot is provided, are the axes, title, and datum clearly defined?

### Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

* Have you thoroughly summarized the entire process you used for this project?
* Were there any interesting aspects of the project?
* Were there any difficult aspects of the project?
* Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

### Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

* Are there further improvements that could be made on the algorithms or techniques you used in this project?
* Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?
* If you used your final solution as the new benchmark, do you think an even better solution exists?

**Before submitting, ask yourself. . .**

* Does the project report you’ve written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your analysis, methods, and results?
* Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?
* Is the code that implements your solution easily readable and properly commented?
* Does the code execute without error and produce results similar to those reported?

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 [↑](#footnote-ref-1)
2. Garson, G.D. 1991. Interpreting neural network connection weights. Artificial Intelligence Expert. 6(4):46–51. [↑](#footnote-ref-2)