**Educational Outcomes for Deaf and Hard-of-Hearing Children**

**In the Denver Public Schools**

Jeff Coady

Data Science Capstone

August, 2019

**Abstract.** This project examined state achievement test reading outcomes for deaf and hard-of hearing children in a large, urban school district. Test results obtained from the Denver Public Schools (DPS) were analyzed to predict Reading scores from demographic information. Data were analyzed via an LSTM model that accounted for 79.77 percent of the variance in reading scores.

**Background and Research Question.**

In 1997, the State of Colorado instituted Universal Newborn Hearing Screening. The rationale was that earlier identification would lead to early intervention so that deaf and hard-of-hearing (DHH) children would not lag behind their peers upon starting school. Without such interventions, DHH kids can

The Colorado Home Intervention Program (CHIP) at CU-Boulder has been tracking DHH children’s development since the mid-1990s. The lab director, Christie Yoshinaga-Itano, was recently awarded CDC grant funding to tie the results of those early language and cognition tests to later school performance. A good measure of school age abilities is the Colorado Student Assessment Program (CSAP) Reading score. It is administered to all students in grades 3-10, and so provides a good measure of educational progress for elementary, middle, and high school students. As a first step toward this goal, the current study seeks to model CSAP reading scores from hearing loss, disability, and demographic information.

Neither the State nor DPS publishes average CSAP scores by grade. Instead, they report the percentage of students in categories ‘Advanced’, ‘Proficient’, ‘Partially Proficient’, and ‘Unsatisfactory’. There is also no information about category cutoffs, or which scores constitute which categories. The most recently published results (2014) show that the distribution for DHH students’ scores are skewed to the low end of the statewide distribution.

|  |  |  |
| --- | --- | --- |
|  | **Statewide** | **DHH** |
| Advanced | 7 | 0 |
| Proficient | 65 | 27 |
| Partial Prof | 18 | 28 |
| Unsatisfactory | 10 | 43 |
| No Score | 0 | 2 |

**Data Collection**. Test scores and demographic information were provided to the CHIP Lab by DPS. All test scores for all DHH children from the 2000-2001 school year to the 2016-2017 school year were contained in an Excel spreadsheet with shape 289456, 14. Each row contained a single test score, along with school year, test, student number, grade, school, and score. A separate demographic worksheet contained information related to student number, school year, gifted/ talented status, free/reduced lunch status, Section 504 status, special education status, and disability status.

**Data Wrangling.** Data wrangling included filling in missing values where possible, removing null values, converting data types, one-hot encoding of binary variables, pivoting and merging tables, and scaling and transforming data for the machine learning task.

CSAP reading scores included many missing values, as families moved into and out of the school district. Further, the State allows three exemptions from the mandatory test: (1) students who move into the school district after October 1, (2) students whose dominant language is not English for the first three years in public schools, and (3) students with disabilities who take an alternate form (CSAPA).

The test data spreadsheet contained 289,456 individual test scores representing data from 643 students. 338 of those students had at least one CSAP Reading score, with an average of 4.37 per student. Data were organized so that predictor variables and a CSAP score in one grade were used to predict the CSAP score in the next grade. This required that data be limited to cases of recorded CSAP scores in consecutive years. That left 833 CSAP scores representing 204 different students.

Predictor variables were:

(1) Student Number, representing within-subject differences

(2) Home Language

(3) Laterality where 0 = unilateral loss, 1 = bilateral loss

(4) Degree of Hearing Loss, where 0 = mild/moderate, 1 = severe/profound/ cochlear implant

(5) Disability, where 0 = hearing loss only, 1 = hearing loss plus other disabilities

(6) Gifted/Talented as a proportion of years students were involved in the program

(7) Free/Reduced Lunch as a proportion of years

(8) Section 504 as a proportion of years

(9) Special Education as a proportion of years

(10) CSAP Reading score at a particular grade

(11) Grade at which that CSAP score was recorded.

The dependent measure to be predicted was the CSAP score in the following grade.

Data were scaled and transformed, and split (.75/.25) into train and test sets.

**Data Visualization.** The following graphs display CSAP scores relative to predictor variables.

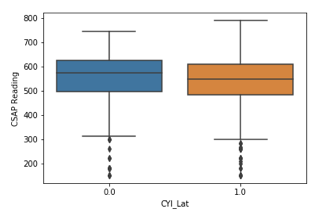
Grade. A boxplot of CSAP scores by grade shows scores increasing by grade, R2=0.017, p < .001.

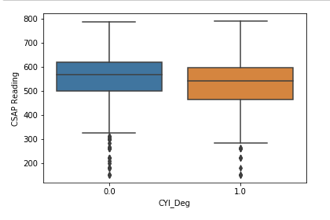
A close up of a clock

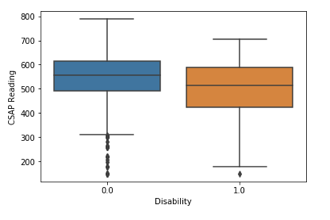
Description automatically generated

A screenshot of a cell phone

Description automatically generatedHome Language. The boxplot of CSAP scores by home language shows that students from English-speaking homes score higher than students from non-English speaking homes, R2=0.037, p < .001.

Laterality. Students with unilateral hearing loss and those with bilateral loss have comparable CSAP Reading scores, t = 1.57, p = .12.

Degree of Hearing Loss. Students with mild-moderate hearing loss have CSAP scores higher than those with severe-profound-CI hearing loss, t = 3.94, p < .001.

Disability. Children with just hearing loss have higher CSAP Reading scores than children with hearing loss and other impairments, t = 4.61, p < .001.

A screenshot of a cell phone

Description automatically generatedGifted/Talented. CSAP Reading scores rose with participation in the Gifted/Talented program, R2=0.12, p < .001.

A screenshot of a cell phone

Description automatically generatedFree/Reduced Lunch. Free/reduced lunch is a proxy for socio-economic status (SES). Children from lower SES back-grounds tend to lag behind their higher-SES peers in both language development and school achieve-ment. In the current study, CSAP scores fell with increasing participation in the free/

reduced lunch program, R2=0.12, p < .001.

A screenshot of a cell phone

Description automatically generatedSection 504. Section 504 is meant to monitor at-risk children who don’t actually qualify for services. For DHH students, the program provides amplification services, and so is a very useful service. CSAP scores rose with participation in Section 504, R2 = 0.10, p < .001.

A screenshot of a cell phone

Description automatically generated

Special Education. CSAP scores fell as participation in special education rose, R2 = 0.12, p < .001.

**Modelling.** The input data were organized as a set of unchanging predictor variables for each row plus the CSAP score and student grade at time(t), all to predict the CSAP score at time(t+1). Since there is a time component to this data, a Long Short-Term Memory (LSTM) model was used. LSTM is a type of Recurrent Neural Network (RNN) in which nodes receive input and emit output, as per usual. But they also include loops that allow information to persist, making them a natural fit for time-series data.

A close up of a clock

Description automatically generated

The LSTM model was implemented though keras and had an input layer of 256 nodes, a hidden layer with 256 nodes, and an output layer with a single node. The input layer had a dropout rate of 0.2, and the hidden layer had a dropout rate of 0.5, temporarily removing 20 percent of nodes from the input layer and 50 percent of nodes from the hidden layer at each iteration. According to Srivastava et al. (2014), this forces nodes in the hidden layer to form more robust connections with nodes from the input layer since they can’t consistently rely on other hidden-layer nodes.

The model was set to run with epochs = 50, each with a batch size = 60, or approxi-mately 10 percent of the training data. The model was also set up for early stopping in the case where the model does not improve over subsequent epochs. Because the model was meant to predict a continuous value, it was technically a regression analysis, and so loss = mean\_squared\_error. Also, the model set the validation\_split = 0.1, where the model was trained with 90 percent of the input data at each epoch, with the remaining 10 percent reserved to adjust the loss parameter for that epoch.

**Model Results.** The most recent version of the model stopped early, after 43 epochs. The graph of the loss shows that loss stabilized by about the fifth epoch. Loss for both the train and test sets didn’t really change after that.

A screenshot of a cell phone

Description automatically generated

Predictions were made from both the train and test data, and those predictions were transformed back into the original scaling. Mean squared error (MSE) was calculated as the squared distance between true and predicted values for both sets. MSE for the training set was R2 = 0.7052, and MSE for the test set was R2 = 0.7977.

A close up of a map

Description automatically generated

The scatterplot above shows the agreement between true and predicted values, with higher agreement indicated by dots closer to the red diagonal. The model performed better with average and higher CSAP scores, but less well with lower scores. It looks like the model was predicting higher values for the extremely low actual values.

A close up of a map

Description automatically generated

The scatterplot of the test values shows a similar pattern, with most dots close to the diagonal indicating good agreement between true and predicted values. Interestingly, there weren’t as many extremely low true values, meaning that they weren’t underrepresented in the train data.

**Conclusion.** After training the model on demographic and previous test score data, it successfully accounted for 79.77 percent of the variance in the test data. I am quite happy and impressed with this result. However, I am curious to know which of the predictor variables accounted for significant portions of the variance in the outcome variable. I plan to follow up this analysis up with a more traditional regression that will delineate the roles of all the input variables.

I also hope to incorporate other predictor variables into these analyses. While this was a great first step, it was limited to just the demographic information, and did not include early childhood test results. I can only imagine that including those variables would only strengthen the model.

**References.**

Li, S. Time Series Analysis, Visualization & Forecasting with LSTM. https://towardsdatascience.com/time-series-analysis-visualization-forecasting-with-lstm-77a905180eba

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15,* 1929-1958.