CS 109A Final Project Report

Sophia Cho, Bach Nguyen, Thu Pham, Jonathan Zhang

December 10, 2021

1 Introduction and Motivation

The COVID-19 pandemic has disrupted education at all levels, constraining educational opportunities for students around the world. As schools closed, teaching was undertaken remotely on digital platforms, making students' access to technology and the internet evermore critical. In this project, we aim to gain deeper insights into the relationship between students' internet connectivity and their success at school, measured by attendance rates, completion rates, and out-of-school rates for primary and secondary schools across the globe. Furthermore, we try to identify the most significant predictors of students' success at school and develop models that best predict their success amid this unprecedented time.

2 Data and EDA

2.1 Description of Data

We used four datasets from UNICEF that look at the proportion of students who are out of school, school attendance rates, school completion rates, and internet connectivity rates by country. These datasets are separated into primary, lower secondary, and upper secondary education levels (pre-primary was available in only two of the datasets, so we dropped it for consistency). Each country's data was collected and divided into categories such as gender, residence (rural or urban), and wealth.

2.2 Data Cleaning

Our first step was to clean the data and make the four Excel files more uniform, then to export the files as CSV files so that they were executable in our code. Specifically, we dropped any columns involving the data source, as these would not be important to our model. Furthermore, we relabelled columns accordingly to avoid confusion in the merged datasets (e.g. each dataset has percentage splits by urban and rural, so we prepended the columns with text to specify which dataset they were coming from). After this cleaning, we decided to merge the datasets, split only by primary, lower secondary, and upper secondary education levels.

When merging the datasets, we noticed that the internet connectivity dataset had less information (i.e. columns) than the other datasets. Specifically, it didn't have splits by gender, and it didn't have the three stratifications for the wealth percentiles, in between poorest and richest (second, middle, and fourth). However, it was consistent with the other datasets in labeling the countries by development level, As a result, we made the decision to not include these extra pieces of information in the wealth stratification across our general model, and to instead use development level for sake of consistency.

Further, we decided to use kNN imputation to deal with missing data. Thus, to make our kNN more well-informed, we imported 7 additional datasets from UNESCO that look at fertility rate, GDP per capita, average life expectancy, mortality rate, population growth, rural population proportions, and total population by country. Within

each dataset, we chose to only look at the 2019 values, since 2019 was the most recent year present across all of the datasets. We merged the datasets, then merged this newly created dataframe in conjunction with the primary, lower secondary, and upper secondary dataframes.

It is also important to note that there were certain countries where imputation was not possible, and this is because they were completely missing data for all the columns. Interestingly, these two countries happened to be USA and Canada, although it is unclear why these two countries were missing a significant amount of data. We decided to drop these two countries altogether from the dataset.

2.3 Exploratory Data Analysis

We looked into the relationship between internet connectivity and out-of-school, attendance, and completion rates. Below are our findings.

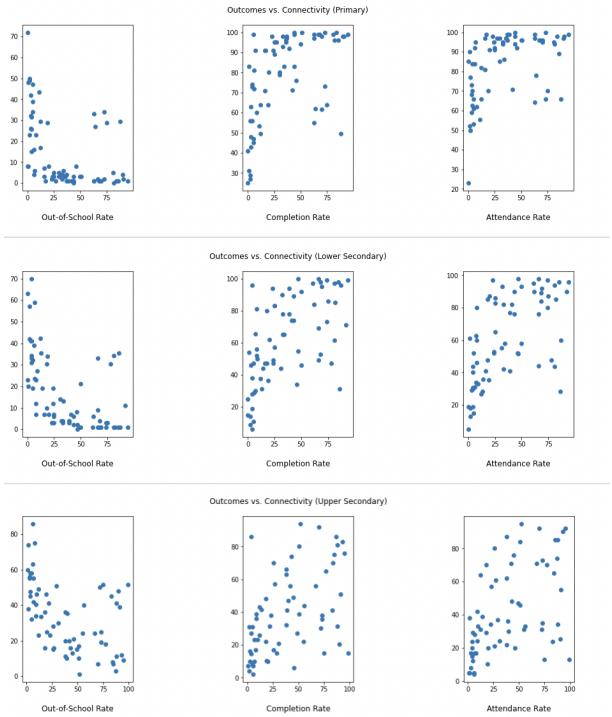

Figure 1: Internet Connectivity vs. Out-of-School, Completion, and Attendance Rates

From the above plots, there seems to be a negative correlation between internet connectivity and out-of-school rates and a positive correlation between internet connectivity and attendance and completion rates.

3 Methods

Our baseline models were linear regression models that predicted attendance (`att_Total`), completion (`comp_Total`), and out-of-school rates (`oos_Total`) for every education level from internet connectivity (`con_Total`). From there, we examined more complex models with more predictors. For every method that we performed, we did so with nine iterations (the three rates we're predicting $\times$ the three education levels).

Our models with no systematic variable selection were:

- Linear regression with one predictor, `con_Total` (a.k.a. the baseline model)

- Linear regression with two predictors, `con_Total` and development level
- Polynomial regression with one predictor, `con_Total`
  - This model was inspired by the fact that our R^2 values for linear regression with one predictor were quite bad, but we were hesitant to immediately build a much more complicated model.
  - To choose the best degree for the polynomial regression, we followed the method outlined in Homework 3 Question 5. That is, we first committed to running 100 bootstraps. The maximum degree (`max_degree`) we considered was 16. For each bootstrap, we executed the following procedure:
    * We used the `cross_validate` feature to perform cross-validation with $k = 10$ for each degree up to `max_degree`.
    * We then selected the best cross-validated degree polynomial regression based on the lowest mean validation MSE, storing that degree to a list.
  - Once we finished our 100 bootstraps, we had a list of 100 best degrees from each iteration.
  - Finally, we generated a histogram to visualize our results, then selected the degree that appeared the most in the list.
- Linear regression with all possible predictors
- Boosting
  - We used the `GradientBoostingRegressor` function for boosting.
  - The process was relatively straightforward: once we had generated a model from the `GradientBoostingRegressor`, we simply fit it on our train data.
  - Finally, to evaluate the boosting performance, we calculated the train and test MSE values.

When we examined more complicated models, however, we had to cut down the number of predictors since there were so many. The sheer number of predictors combined with the complexity of the models overwhelmed our computers with with degrees as small as 2. We used feature importance from random forest as a method for variable selection. A dictionary object was setup where each key refers to each feature and each value refers to the number of times this feature was considered significant. 3 of such objects were initialized, each one for each education level. It should be noted that we had three separate sets of predictors: those to predict attendance rate, those to predict completion rate, and those to predict out-of-school rate, and this is reflected in the keys of the dictionaries setup from initialization. The reason for this is to prevent highly correlated predictors. Then for each of the 9 iterations, we trained and fit the `RandomForestRegressor` instance and summed up all the importance scores for those predictors where the feature importance was greater than 0. Once we were done with collecting the counts, for each education level we picked the top 10 predictors. We specifically picked 10 predictors due to limitations in the hardware used to train models. We plotted the importance of all of the predictors used for each response variable below.

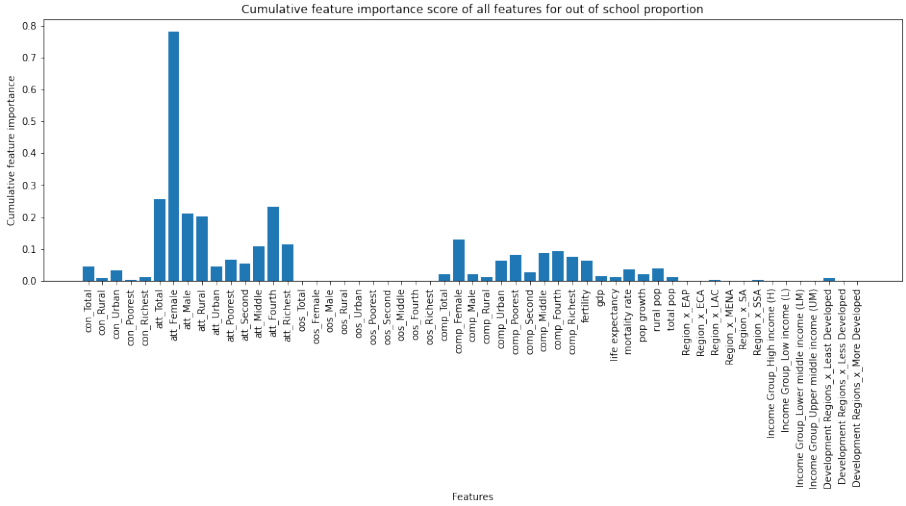

Figure 2: Cumulative Importance Scores for Out-of-School Rate Predictors

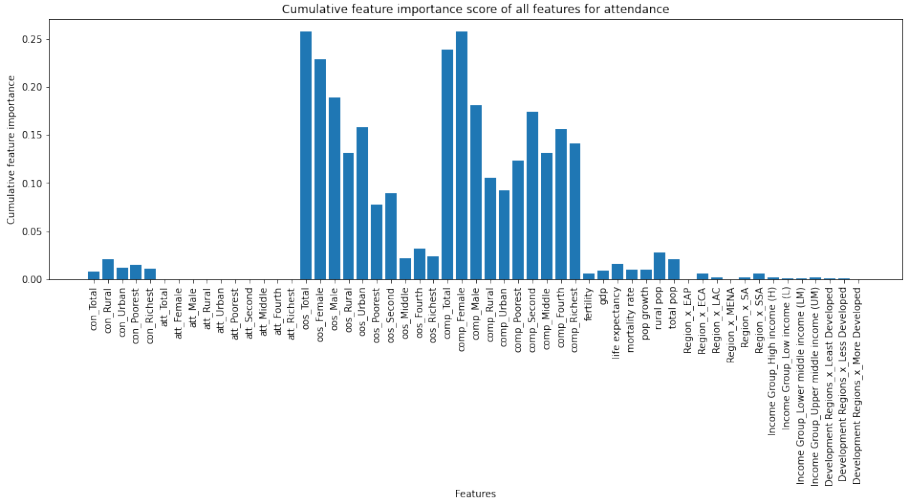

Figure 3: Cumulative Importance Scores for Attendance Rate Predictors

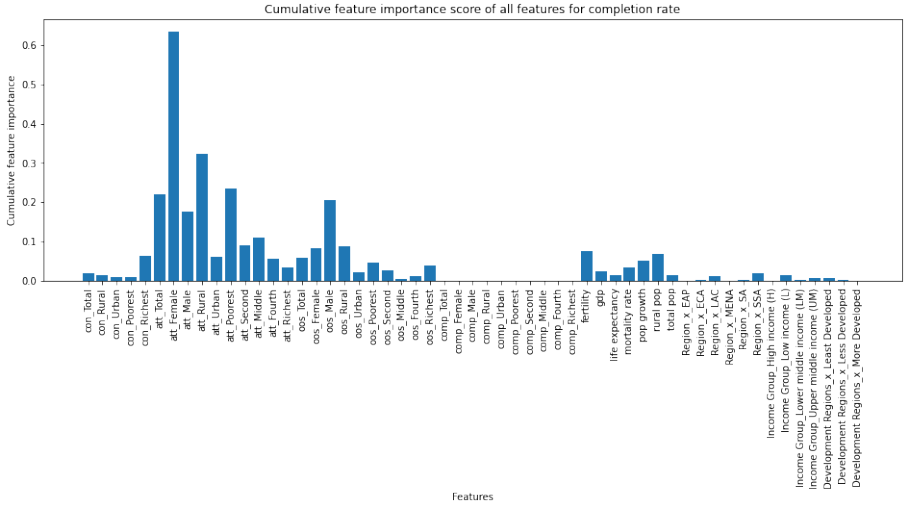

Figure 4: Cumulative Importance Scores for Completion Rate Predictors

Once we selected the predictors to use in our models, we proceeded with the following:

- Linear regression with all selected predictors
- A model with all possible two-way interaction terms from selected variables, using LASSO regularization
  - To generate all possible two-way interaction terms, we used the `PolynomialFeatures` function, with `interaction_only` set to `True` and `degree` set to 2.
  - From there, we tuned our hyperparameter α for the LASSO regression, using cross validation, selecting values of α from the list $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}]$.
  - After ten-fold cross validation, we found the best value of α for the nine versions of the model: predicting the attendance rate, completion rate, and out-of-school rate for education levels of primary, lower secondary, and upper secondary.
  - Taking these nine individual best α values, we computed the train and test MSE values for each of the nine models with LASSO regularization and interaction terms.
- Polynomial regression, using LASSO regularization
  - This model was inspired by our previous model, which just considered all possible two-way interaction terms between the selected predictors. We wanted to add more complexity by also considering the pure polynomial terms (i.e. the squared terms).
  - Because this was a direct expansion of the interaction term model, we used degree 2 but set our `interaction_only` term to `False`.
  - For every possible α from the list $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}]$, we used cross-validation with ten folds.
  - After cross-validating, we found the best value of α for each of the nine versions of the model: predicting the attendance rate, completion rate, and out-of-school rate for education levels of primary, lower secondary, and upper secondary.
  - Taking these nine individual best α values, we computed the train and test MSE values for each of the nine models with LASSO regularization and interaction terms.

Finally, to compare all of the models, we calculated their train and test MSE values, which we lay out in the following Results section.

4 Results

	pri oos	pri att	pri comp	lsec oos	lsec att	lsec comp	usec oos	usec att	usec comp
one pred	219.800	221.656	346.712	212.718	457.625	475.858	284.087	525.842	484.854
two preds	164.293	178.693	242.217	152.518	346.642	351.940	223.864	417.725	392.420
poly one pred	144.040	154.720	235.567	141.587	378.600	418.999	205.546	479.437	484.854
OLS all preds	15.993	10.254	44.569	48.119	74.231	86.398	55.710	50.124	78.513
OLS selected preds	17.651	19.076	89.830	63.019	183.944	154.300	100.649	206.625	145.951
poly LASSO	7.321	18.921	62.440	34.780	96.893	69.461	55.545	66.064	50.445
interaction LASSO	7.585	20.200	61.873	35.236	122.467	77.979	61.207	76.653	55.690
boosting	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 5: Test MSE Values for All Models, Education Levels, and Rates

	pri oos	pri att	pri comp	lsec oos	lsec att	lsec comp	usec oos	usec att	usec comp
one pred	-134.113	-104.503	3.098	-114.757	123.945	42.105	234.495	223.007	575.542
two preds	-84.222	-105.866	-33.627	18.694	30.654	57.275	447.981	293.273	728.177
poly one pred	-58.393	-29.188	166.497	-32.721	32.826	-10.271	52.198	-243.231	-7.447
OLS all preds	72.526	36.143	214.677	45.673	359.153	63.251	464.592	244.956	223.310
OLS selected preds	5.986	14.716	79.325	121.445	284.211	12.074	195.354	151.628	49.002
poly LASSO	79.580	176.771	168.329	1298.826	4144.081	4966.535	282.144	576.519	295.457
interaction LASSO	70.252	53.983	177.341	167.667	936.738	848.244	286.237	690.475	1060.879
boosting	16.245	32.787	217.866	148.850	291.309	150.726	244.673	285.459	142.015

Figure 6: Difference in MSE Values for All Models, Education Levels, and Rates

5 Discussions and Conclusion

5.1 Discussions

To evaluate the overall performance of each model, we looked at the models that yielded the lowest test MSE for each of the nine iterations (predicting out-of-school, attendance, and completion rates for the three different school levels). The results are listed below:

- The best model to predict **out-of-school rates** at the **primary** level is boosting, with a test MSE of 16.245.
- The best model to predict **attendance rates** at the **primary** level is boosting, with a test MSE of 32.787.
- The best model to predict **completion rates** at the **primary** level is a multiple linear regression with the predictors selected by random forest, with a test MSE of 169.155.
- The best model to predict **out-of-school rates** at the **lower secondary** level is a multiple linear regression with all possible predictors, with a test MSE of 90.875.
- The best model to predict **attendance rates** at the **lower secondary** level is boosting, with a test MSE of 291.309.
- The best model to predict **completion rates** at the **lower secondary** level is a multiple linear regression with all possible predictors, with a test MSE of 150.533. *(Of note: the boosting method was a close second, with a test MSE of 150.726.)*
- The best model to predict **out-of-school rates** at the **upper secondary** level is boosting, with a test MSE of 244.673.
- The best model to predict **completion rates** at the **upper secondary** level is boosting, with a test MSE of 285.459.
- The best model to predict **attendance rates** at the **upper secondary** level is boosting, with a test MSE of 142.015.

Overall, we saw that boosting overwhelmingly performed the best, and multiple linear regression with all possible predictors had the second best performance with the lowest test MSE for two iterations of our predictions. It makes sense that boosting performed as well as it did, as it reduces both bias and variance, by combining a set of weak learners to create one strong one. Boosting also does not rely on any distributional assumptions (unlike linear regression, for example), so we do not need to check any assumptions. Furthermore, the lack of assumptions (and thus its inherent robustness to any deviations from Normality, linear relationships, etc.) required for this model may also explain its high performance.

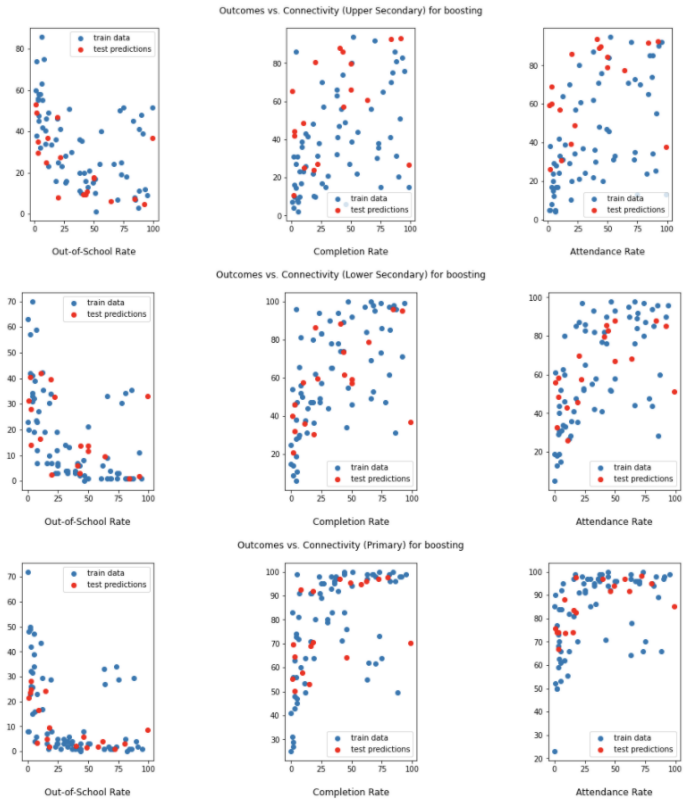

Figure 7: Boosting Prediction Plots

Once we had chosen our best model, we returned to our framing question about the relationship between connectivity and our three predictors. As seen above, we plotted the boosting predictions (holding all other predictors constant) with connectivity as our independent variable. As seen in our EDA, we observed a negative correlation between out-of-school rates and connectivity, and a positive one between attendance and completion rates and connectivity. Though we found similar results in terms of the correlation, we now have a more powerful model to predict these response variables.

Furthermore, we evaluated each model's tendency to overfit by examining the difference in train and test MSE values. We reported the models that are most likely to overfit and least likely to overfit, for the nine different iterations. We had several instances where our test MSE was lower than our train MSE. Because overfitting is measured by how much larger the test MSE is than the train MSE, we ignored any negative differences of test MSE minus train MSE. We discuss the results below:

- For predicting **out-of-school rates** at the **primary** level, the model least likely to overfit is the multiple linear regression on the predictors selected by the random forest. The model most likely to overfit is the multiple linear regression on all possible predictors.
- For predicting **attendance rates** at the **primary** level, the model least likely to overfit is the linear regression with connectivity as a single predictor. The model most likely to overfit is the polynomial terms with LASSO regularization.
- For predicting **completion rates** at the **primary** level, the model least likely to overfit is the linear regression with connectivity as a single predictor. The model most likely to overfit is boosting.
- For predicting **out-of-school rates** at the **lower secondary** level, the model least likely to overfit is the linear regression on two predictors, connectivity and

development level. The model most likely to overfit is the polynomial regression with LASSO regularization.

- For predicting **attendance rates** at the **lower secondary** level, the model least likely to overfit is the linear regression on two predictors, connectivity and development level. The model most likely to overfit is the polynomial regression with LASSO regularization.
- For predicting **completion rates** at the **lower secondary** level, the model least likely to overfit is the linear regression with the predictors selected by the random forest. The model most likely to overfit is the polynomial regression with LASSO regularization.
- For predicting **out-of-school rates** at the **upper secondary** level, the model least likely to overfit is the polynomial regression on one predictor, connectivity. The model most likely to overfit is the multiple linear regression on all possible predictors.
- For predicting **attendance rates** at the **upper secondary** level, the model least likely to overfit is the multiple linear regression with predictors selected by the random forest. The model most likely to overfit is the two-way interaction terms with LASSO regularization.
- For predicting **completion rates** at the **upper secondary** level, the model least likely to overfit is the multiple linear regression with predictors selected by the random forest. The model most likely to overfit is the two-way interaction terms with LASSO regularization.

Overall, the models that are least likely to overfit (in order) are the multiple linear regression on the predictors selected by random forest, for four of the predictions; linear regression on the single predictor connectivity, for three of the predictions; and linear regression on connectivity and development level, for two of the predictions. These results make sense, as they are the simplest in terms of structure, being linear regressions. Furthermore, the number of predictors are limited in some way (as a large number of predictors can lead to overfitting), whether it's by variable selection or a personal choice to only use one or two predictors.

Furthermore, the models that are most likely to overfit (in order) are the polynomial regression on selected variables with LASSO regularization, for four of the predictions; the multiple linear regression on all predictors, for two of the predictions; and the two-way interactions between the selection predictors with LASSO regularization, for two of the predictions. These results align with our understanding of the models. The polynomial regression overfits the most because it is so complex, and even though we capped the degree at 2 and used LASSO regularization to control for overfitting after the fact, the high degree still causes us to overfit on the train data. The two-way interactions plus LASSO regularization overfits for a similar reason; even though we both selected variables ahead of time and then used LASSO regularization, the model is so complex that it tends to overfit. Finally, it makes sense that the multiple linear regression with all predictors overfits a nontrivial amount of times, as we know that having a high amount of predictors will lead to overfitting.

Overall, if we take the two results together, we see that there is very little overlap with models that perform best (lowest test MSE) and models that are least likely to overfit (lowest difference between test and train MSEs, for positive differences only). Thus, we see a trade-off that is common in data science: balancing models that perform well with those that tend to overfit on train data.

5.2 Limitations

There were three main limitations we faced in the execution of our project. First, there was a fair amount of missing data that we had to impute, limiting the accuracy of

our models. Second, since we had so many predictors, we were not able to use all of them to predict out-of-school, attendance, and completion rates. Instead, we had to do variable selection through random forest. Finally, we noticed how slowly some of our code ran. Although this was to be expected for some of the most complex models, we were surprised that our code was so slow for even degree 2 (a relatively low degree) on our predictors selected by random forest. Even though this model overfit anyways, we still consider this a limitation moving forward.

5.3 Future Work

In our project, we decided to use LASSO as our method of regularization. We did so because we hypothesized that we had a small number of strong predictors, instead of a large number of weak ones. If we had more time (given that we have nine of each model, and thus were a little tight on time and bandwidth), we certainly would have explored the results of LASSO regularization as well. We would then compare the two's performance based on the results of our variable selection.

In addition, we would also like to investigate any instances where our test MSE was less than our train MSE – perhaps this would require more cross validation. We also notice that some models are very prone to overfitting, even after controlling for that to some degree through regularization and variable selection. Thus, it would be interesting to investigate further methods of overfitting to make use of certain models. Furthermore, we would like to examine other ways to evaluate our models besides MSE, especially if they yield different rankings. Finally, we'd like to merge other datasets with the ones provided from UNICEF, both to help fill in any missing data points or add information on other dimensions.

6 Acknowledgements

We would like to thank the entire CS 109A teaching staff for leading us through a fruitful and knowledge-filled semester.