Prediction of the Educational Enrollment of Countries

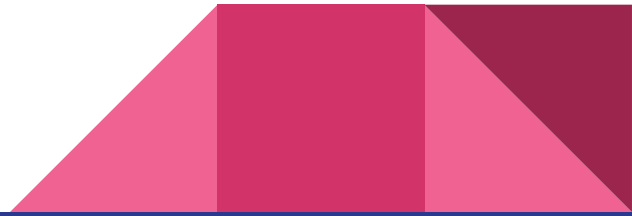
Chance Peterson
Sourav Ukil

Problem Statement

Predict education data based on more readily available data about countries.

We wanted to predict net educational enrollment.

We used factors such as country GDP per capita, GNI per capita, population, among others.



Project Purpose & Current Approaches

There are many countries struggling with education around the world.

Being able to identify potential countries in need of assistance is valuable.

Many current methods of classifying countries requires testing and surveying school children directly.

This method is costly and difficult to implement.

Solution Approach

We approached this problem as a regression problem

We wanted to find a line that would fit the data well and allow us to make predictions using new data.

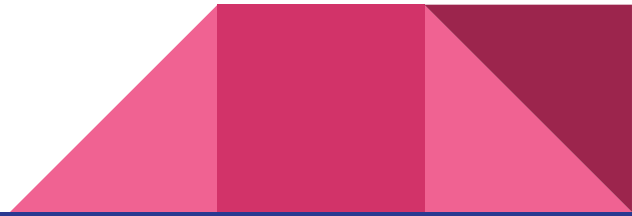
We used Python 3.0 with toolkits including scikit-learn and pandas.

These tools helped us preprocess our data and implement our regression models.

Education Dataset

We have taken data from
https://datacatalog.worldbank.org/dataset/education-statistics

The World Bank EdStats All Indicator Query holds over 4,000 internationally comparable indicators that describe education progression, completion, literacy, population, and expenditures.



Dataset

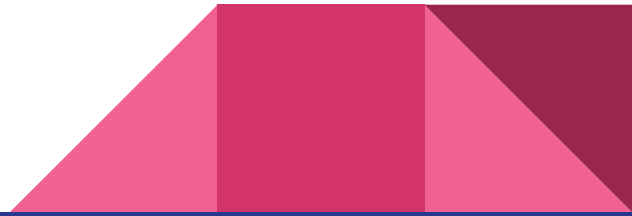
The dataset had many missing entries.

We dropped the samples with missing response variables.

We split the dataset into two sets: training and test.

To fill in the missing values, we tried two techniques:

- Imputing the mean of the features
- Matrix completion using matrix factorization



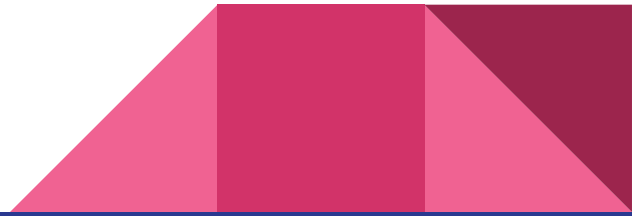
Matrix Completion

Correlation matrix results:

We found that GDP per capita, GNI per capita and Internet Users/100 were correlated with each other, but not with the others.

All the population variables were correlated

We split the features into 2 sets, per above.

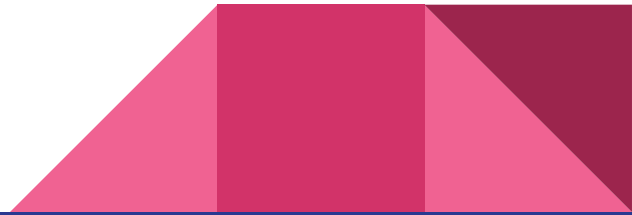


Matrix Completion

Let Z be an m by n matrix. We want to find matrix X (m by l) and matrix Y (l by n) such that X * Y is very close to Z everywhere that Z has a known value. To do this, we used stochastic gradient descent to optimize the objective function.

Objective function: Least squares error between each known value of Z with the corresponding entry of the matrix X * Y.

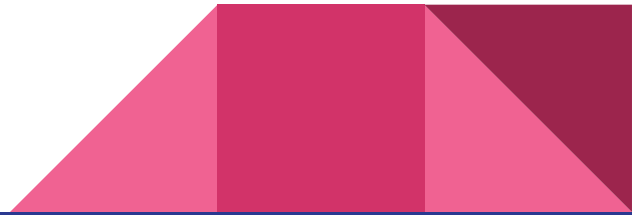
Once we find an X and Y that result in a satisfactory error rate, we can use the entries of X * Y to estimate the missing entries of Z.



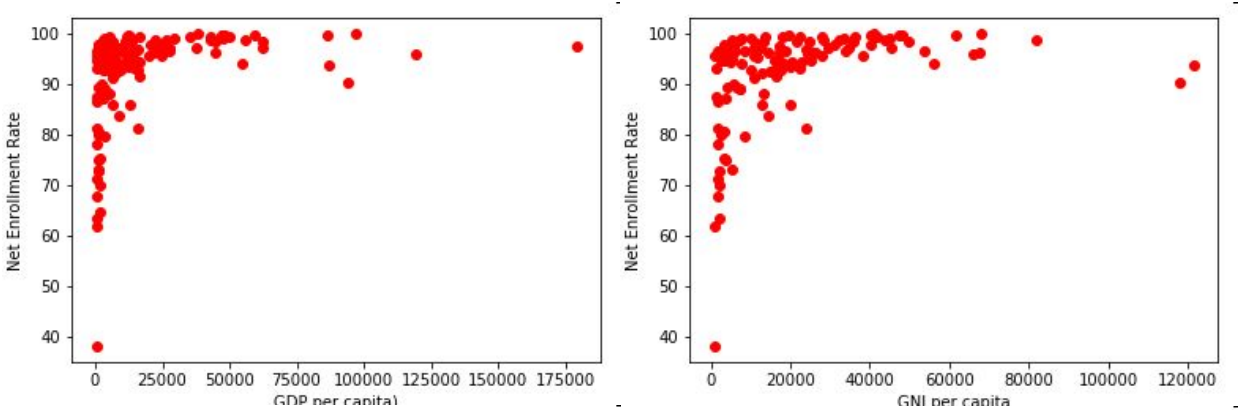
Processed Dataset

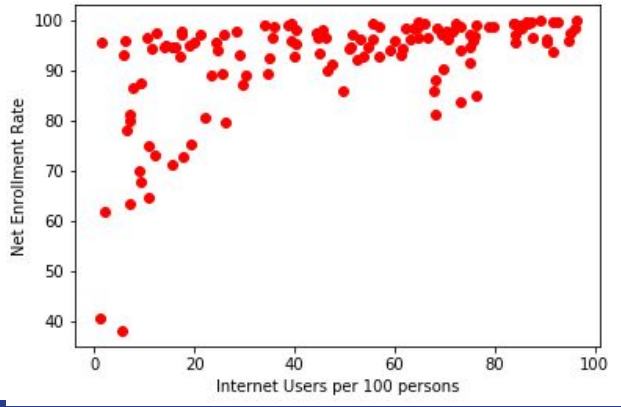
We now have two datasets with missing entries filled in using different methods.

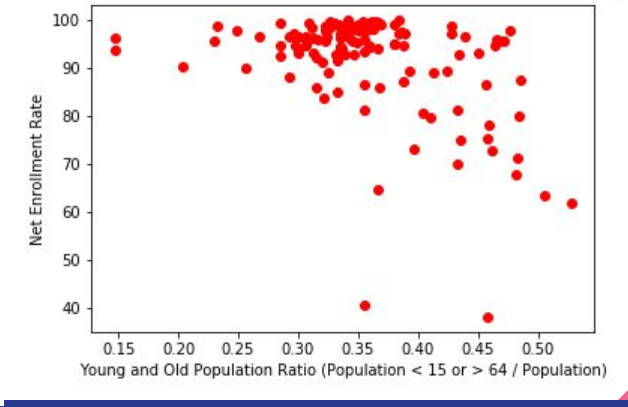
Now, to visualize the relationships between single features and the response.



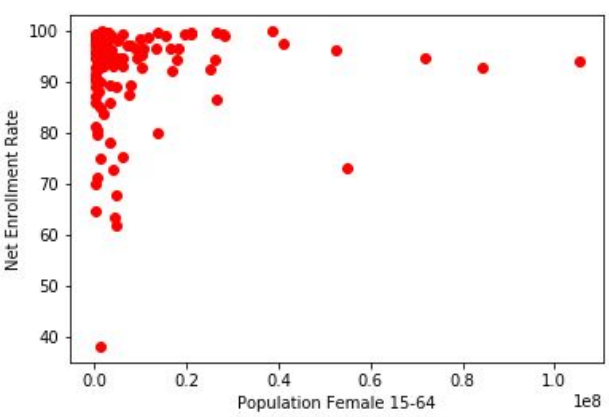
Visualizing Some Features

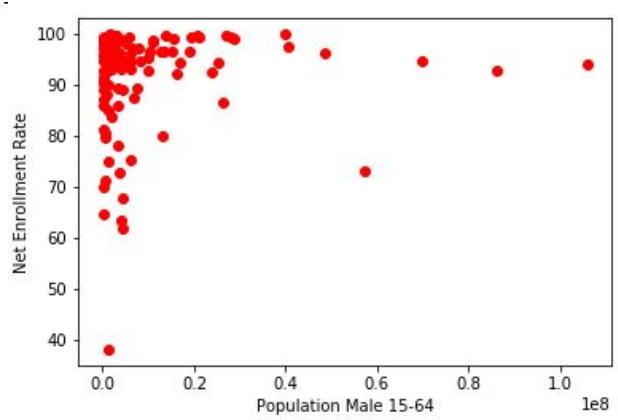


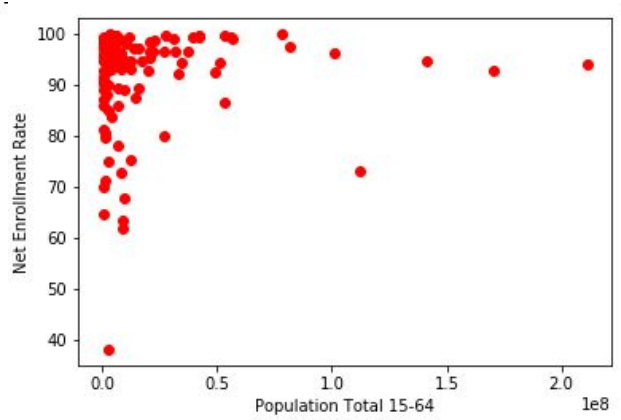




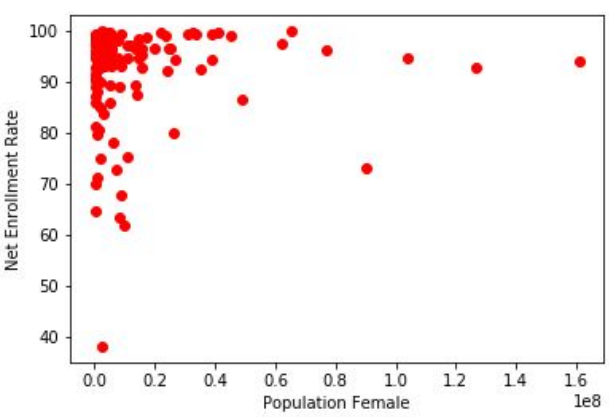
Visualizing Population Features

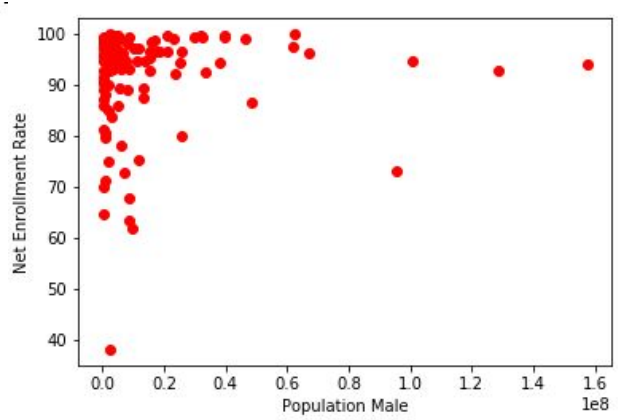


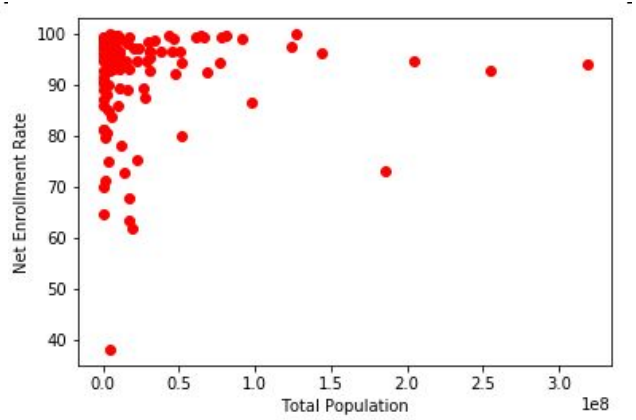




Visualizing Population Features





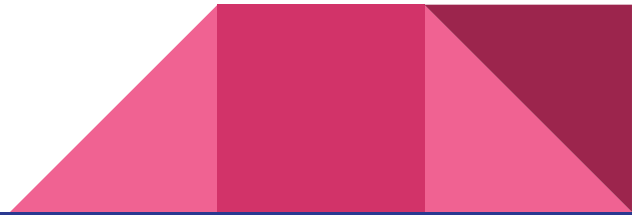


Optimization Modeling Approach

We will go with the Least Squares Method to find a line of best fit. We used the closed form solution for to solve the Linear Regression Problem.

Objective Function: Least squares error rate, which measures the distance between the fitted curve and the data.

Constraints: None



Modeling

First we used simple linear regression using features:

GDP per capita, GNI per capita, Internet Users per 100, Young+Senior Population Ratio

r^2	Train	Test
Mean filled dataset	.268	.337
Matrix completion dataset	.019	---*

*represents an arbitrarily bad score (often comes out as a large negative number)

Modeling

Since the data looked curved between many features and the response we decided to try polynomial regression.

Example for regression with 2 features (x and y):

Create new features x^2, x, y^2, y, xy, b

Perform linear regression on these features.

This gives us the coefficients of each quadratic term.

This problem grows large with more variables and higher powers.

Modeling

Results of quadratic regression:

r^2	Train	Test
Mean filled dataset	.435	.534
Matrix completion dataset	.083	---

Modeling

Trying the same with slightly different features

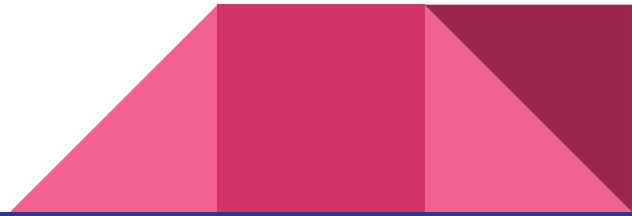
GDP per capita, GNI per capita, Internet Users/100, Female Population 15+, Female Population, Male Population 15+, Male Population, Total Population 15+, Total Population

r^2	Train	Test
Mean filled dataset	.330	.374
Matrix completion dataset	.137	.202

Modeling

Results of quadratic regression:

r^2	Train	Test
Mean filled dataset	---	.755
Matrix completion dataset	.316	.325



Results

The quadratic regression using the original set of mean imputed features performed the best

r^2 train	r^2 test
.435	.534

Results

Let GDP per capita = x, GNI per capita = y, Internet Users per 100 = z,
Young population ratio = a

The equation for predicting net enrollment:

$$-2.33e\text{-}4\ x + 7.38e\text{-}9\ xy + -9.74e\text{-}7\ xz + 7.23e\text{-}4\ xa + -3.95e\text{-}9\ x^2 + -5.69e\text{-}4\ y +$$

$$+ 1.70\ yz + -1.79e\text{-}4\ ya + -7.73e\text{-}9\ y^2 + 9.98e\text{-}2\ z + 1.40\ za + -8.26e\text{-}3\ z^2 +$$

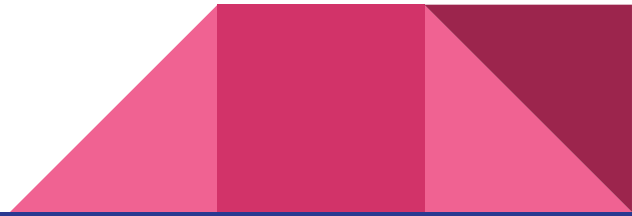
$$+ -9.20e1\ a + -2.27e\text{-}2\ a^2$$

Discussion

Using our best result, we obtained an r^2 value of .435 for the training set and .534 for the test set.

While these results are not optimal, they are expected considering the lack of data for training and testing. It does however show potential for this approach with the use of more complete data and use of other features.

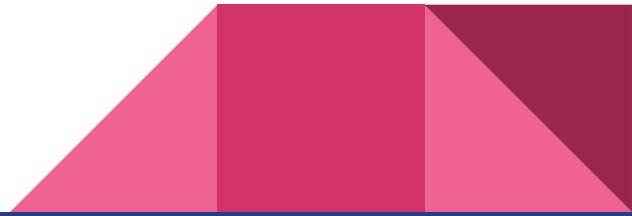
Scikit-learn in Python allowed us to efficiently and quickly implement and test our models.



Impact of the Solution

Being able to predict where educational help is most needed without costly surveying would help expedite the process.

Further studying the relationships between the features and the response can present ideas for possible causes of higher or lower educational enrollment.



Thank You