Analyzing the role of women's rights, income inequality, and education enrollment on a country's Human Development Index (HDI) using multiple linear regression

Contributors:
Milo J. Eirew, *Department of Electrical Engineering (prospective) at Stanford University*
Claudia M. Kania, *Departments of Political Science and Computer Science at Stanford University*
Princess S. Vongchanh, *Symbolic Systems Program at Stanford University*

4.5 Pages + References + Cover Page
Project Poster: https://github.com/pvongchanh12/cs129project/blob/main/Final%20Project%20Poster.pdf
Github Repository: https://github.com/pvongchanh12/cs129project
Video presentation: https://youtu.be/kdPshEonbvE

Abstract (1 paragraph):

In our final project, we sought to create a regression model to predict the HDI of a country given different metrics relating to women's rights, economic inequality, and education. Finding these relationships would be helpful for the field of global development, but the model is only an indicator: it is only tip of the iceberg regarding the actual global factors at play. We approached the problem using multiple regression with polynomial features. To do this, we implemented regularized linear regression normal equations, since the dataset is not prohibitively large. We tuned lambda and the degree of the polynomial features. To tune these hyperparameters, we selected a range of lamba, and a range of polynomial degrees, and trained a regression model on each combination, allowing us to find the ideal value of each. Using these tuned hyperparameters, we were able to create an extremely accurate model, predicting HDI scores very close to the ground truth.

Introduction

Stanford Professor and Researcher Andrew Ng notes that, "artificial intelligence is the new electricity." Machine learning, in particular, has the ability to transform our understanding of the world on various scales and through the analysis of seemingly disparate relationships. Our project attempts to do this by exploring the capabilities of machine learning on human development data.

This research has two key goals. The first is to assess the performance of a multiple linear regression model in predicting the Human Development Index (HDI) of nations around the world by running experiments with different versions of the algorithm. Concretely, we seek to identify which parts of a linear regression algorithm are best to tune for our relatively small dataset, what degree of polynomials produce the lowest error without overfitting, the impact of regularization, and which value for the regularization constant lambda λ gives the best prediction.

The second goal is to identify which of the three following areas best correlates to a country's HDI: women's rights, income inequality, and educational enrollment. The HDI aims to evaluate the socio-economic status of nationals around the world, highlighting people as the ultimate contributors to a country's development (dependent on health and longevity, knowledge, and standards of living). It was created by the United Nations Development Programme, and is calculated in each year's Human Development Report (HDR), in order to assess the impact of federal decisions and social trends on a large scale. Via data manipulation, the HDI is comparable to at least two hundred other data factors, including those upon which this research is based. Our process reverse engineers the HDI's intended purpose, and demonstrates how the HDR can be utilized in practice beyond assumptive reasoning. We believe the HDI should be closely monitored each year and used as a reliable indicator for the development of nations such that the gap between those with high and low indexes does not grow exponentially.

We recognized each of these three areas to be rapidly developing themselves and attempt to argue via this paper that they are worth dedicating significant attention to, both by governments and individuals. We do not claim these are the most significant factors, nor are they the only ones worth analyzing. The aspects of both singular and several interdependent social systems are difficult to delineate; we recognize that the features we analyze are interdependent, and proceed with an understanding that much effort is required to alter even one of these features, nonetheless several of them at once. Hopefully, this research can be utilized by computational social scientists for a range of applications, and inspires stakeholders to engage in changemaking efforts to develop their respective countries.

All of our data and code is available for reference in a public repository on GitHub (https://github.com/pvongchanh12/cs129project).

Related Work

Akin et al. [1] predicted countries' HDI using health indicators and a tree based regression model across a range of four years. The four models they tested–decision tree, Random Forest, extreme gradient boosting, and regular gradient boosting–were each highly effective, with the final method producing the greatest accuracy. This report contributed to our initial choice of a gradient boosted linear regression model which calculates the difference between our model's prediction and the true value; our new model does not utilize this approach (see methods for more). This is also how we calculated our error. The downside to evaluating performance across a range of years as they did was the inability to consider factors independent to a country, such as whether or not they experienced recession or war.

Mulyani et al. [2] conducted research on the relationship between economic growth and income distribution in Indonesia. This research applies mean regression, quantile regression (a robust technique which allows for a holistic review of a relationship), and smoothing splines regression for higher dimensional data. One major limitation of applying the latter two methods is that they are computationally expensive, and require a significant amount of manual labor to differentiate between the efficiency of continuous and discrete functions for their dataset. The size and complexity of our dataset does not allow for an effective application of the two latter methods, but we found their use of mean regression to be beneficial as we proceed with understanding how dependent variables may be considered independently from its coinciding data, in addition to their impact of the coinciding data.

Khan et al. [3] similarly evaluated the ability of multiple machine learning techniques to forecast the HDI of all 189 U.N. member states (at the time of publication) up to the year 2025, with an emphasized comparison to South Asian countries and Bangladesh. We replicated their method for preprocessing data by scaling it in order to improve the model's performance. Their research similarly reverse engineers the intended purpose of the HDI in order to correlate particular features to it, and we found this an effective and thorough approach for analysis. While we found this approach good, their experiment which reverses the inputs and outputs (using the HDI to predict a certain feature) demonstrates extreme accuracy; this suggests that switching the inputs and outputs does not appropriately account for other dependent variables. This highlights the flaws of machine learning, and we decided not to test it.

Dataset

This research project utilizes a custom data set we preprocessed, which compiles percentages of ten different areas within the three aforementioned factors from the United Nations Development Programme's Data Center. It includes data points for 100 member states in the year 2018; if data was unavailable for this year, we used the average of data for the years 2017 and 2019 if both were available and either of the years if only one was available. We selected the year 2018 because we sought to experiment on the reliability of our machine learning model and the HDI without needing to consider the potential for volatile change as a result of the COVID-19 pandemic. Countries which significantly lacked reported data (three or more years) were excluded from the dataset because related work demonstrates that estimating null values (with little to work off of) is not effective. The model may have performed worse had we kept all 189 member states with long term history in the HDRs.

The input of the algorithm includes the following features: unemployment rate (female to male) [4], share of seats in Parliament held by women [5], gender inequality index (GII) [6], inequality in education [7], inequality in income [8], income share held by the richest 10% [9], expected years of schooling [10], education index [11], gross enrollment ratio (secondary education) [12], gross enrollment ratio (tertiary education) [13]. We use a multiple linear regression model to output a predicted HDI [14] based on individual features. There were a total of 100 examples, with training/validation/test sets split 60/20/20.

In order to increase the speed of gradient descent, our features were scaled with the following algorithm:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Methods

We use a multiple linear regression model with polynomial features and regularization. Polynomial features are appropriate because our goal is a regression task: we want to predict HDI given our training features. Linear regression allows us to find the values of theta that fit our the closest. If a feature has a higher theta associated with it, then that feature has a stronger correlation with the output. Although we started off with ten raw data features, our final model has 40 due to the inclusion of higher-degree polynomial features. This allows for greater model complexity, and better performance overall in comparison to the model submitted for the project milestone. In our milestone, we implemented linear regression with quadratic features. With testing (which we discuss in detail in the next section), we discovered that the best fit for the testing set was actually polynomials with a degree of 4. While there is a significant difference in complexity, there is also a risk with implementing a more complex model: overfitting. This requires us to train the model so that it very closely fits the training data, but not so closely that it focuses on the training data; this scenario would not be an accurate representation of the dataset as a whole. In order to prevent overfitting, we implement regularization. Regularization penalizes the usage of more complex features in the cost function: The model is incentivised to represent the model using less complex terms:

$$J(\theta) = (\frac{1}{2m})((X\theta - Y)^T * (X\theta - Y) + \frac{\lambda}{2m} * (\theta^T \theta))$$

A higher regularization constant λ means the higher-order polynomial terms are assigned a lower theta and it decreases overfitting. Between selecting the degree of the polynomial and selecting the regularization parameter, we have two hyperparameters to tune.

Since our dataset is on the small side with 100 examples and 40 features, using normal equations is an appropriate approach because the matrix inverse function will not take a prohibitively long time. Fortunately, we know how to derive the normal equations of regularized linear regression from PSET 1:

* Formulae are from lecture slides or PSET derivations

$$J(\theta) = (\frac{1}{2m})((X\theta - Y)^T * (X\theta - Y) + \frac{\lambda}{2m} * (\theta^T \theta))$$

$$\nabla J(\theta) = (\frac{1}{2*m})[(\theta^T X^T)(X\theta) - 2\theta^T X^T Y + Y^T Y + \lambda * \theta^T * \theta]$$

* We want to find the θ values that result in the smallest possible cost, in other words the minimum, where the gradient of the curve is equal to zero.

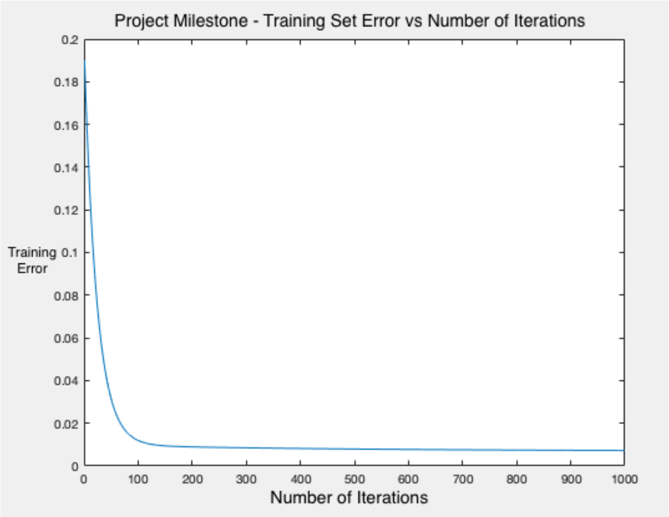
$$(\frac{1}{2*m})[(\theta^T X^T)(X\theta) - 2\theta^T X^T Y + Y^T Y + \lambda * \theta^T * \theta] = 0$$

$$\therefore \theta = [X^T X + \lambda I]^{-1} X^T Y$$

Using the normal equations of regularized linear regression means that we don't need to tune the hyperparameters of step size or iterations; instead, our model converges to the minimum cost, as long as the matrix $[X^T X + \lambda]$ is invertible. If it was uninvertible (a rare case), we would have to change our dataset (such as by adding additional features).

All computations done in our model are vectorized. Rather than computing operations example by example with sums, the computations are done simultaneously.

Using normal equations was the next step from our milestone, implementing gradient descent. We plotted our testing cost over iterations, and noticed that it converged in under 150 iterations. From this, we know gradient descent itself wasn't introducing error:



In our original version of the model with gradient descent, we computed the gradient at each iteration and took a single step towards the minimum. With each iteration, the cost function (= $\frac{1}{2m}\,[h(\theta) - y]^2$) should decrease.

This required use of the vectorized formula for the gradient of the cost function with respect to theta:

$$\nabla J(\theta) = X' * [h(\theta) - y]/m$$
$$= X' * [X\theta - y]/m$$

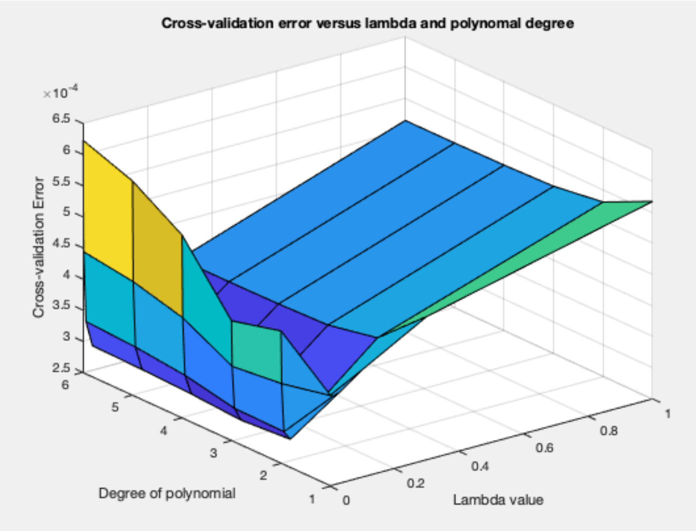
We would then take steps of size alpha α (a hyperparameter we had to find by experimentation) until the cost function converged. This requires identifying another hyperparameter, the number of iterations for our model to perform. Switching to normal equations meant that **we didn't need to tune the step size or number of iterations** in our model. Our new model has prevented additional work and faulty tuning of hyperparameters. Had the step size been too small, the model might not have converged fast enough. Similarly, the number of iterations would have been too small. If the step size was significantly large, we ran the risk of the model overstepping the minimum and actually increasing the cost after every interaction.

Complexity proved to be a very significant factor in improving our model. To properly reflect the complexity of our data, it was important to us that our model was complex in itself. Our final model resulted in polynomial degrees of 3 and 4 performing best, with an ideal regularization parameter of 0.03. This means our model wasn't necessarily limited in complexity; when it was given the option to have higher degree polynomials/lower regularization, its prediction error did not increase. Our research of related work demonstrates that other models could have been more promising to achieve our goals, such as a neural network.

Experiments and Results / Discussion

The two hyperparameters that require tuning in our new model were the regularization constant and the degree of the polynomial (model complexity). Increasing the degree of these parameters increases overfitting of the model, while increasing the regularization parameter decreases model overfitting. Finding a combination of these hyperparameters that struck a balance between overfitting and underfitting the data was key to training a low-error regression model.

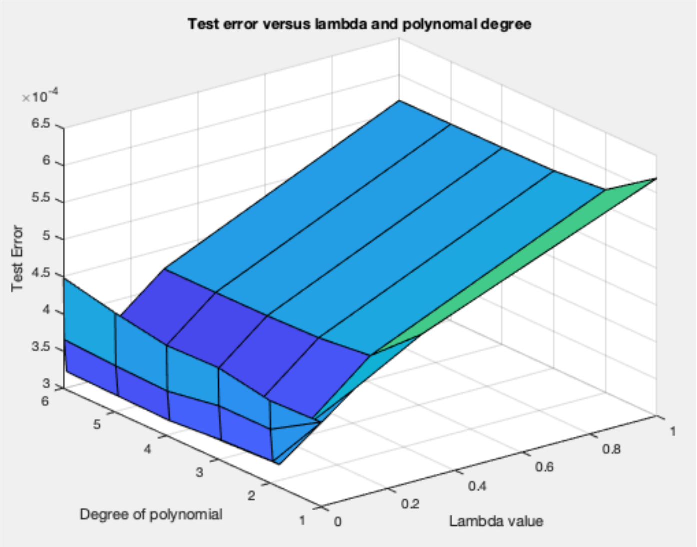
We trained 36 models in total, using different combinations of six values of lambda (ranging from 0.0001 to 1) and six different polynomial degrees (ranging from 1 to 10). For each combination of lambda and polynomial degree, we trained a new model, and evaluated the error of the cross-validation set to determine the ideal values for our hyperparameters. From here, we visualized a 3-dimensional plot of the training and cross-validation (CV) error compared to the lambda value and the degree. We also analyzed it numerically. The results are displayed below, and confirm that our models are working successfully. Since we aren't using gradient descent, we can't visualize the training error over time to confirm the function is working as expected. Instead, this plot tells us that our hyperparameters affect our model as expected.



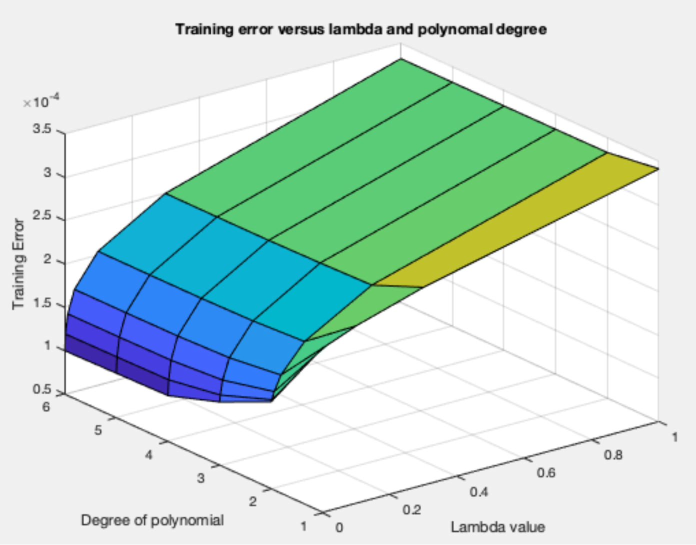
* Lighter color means higher error

Starting with the plot of our CV set versus degree and lambda: we observe a 'scoop shape in the plot.' Error is highest when the polynomial degree is extremely high and regularization is low; this means that the model **overfits** the training data. Error is also high when the polynomial degree is low, and regularization is high; here, the model **underfits** the data. There is a minimum in the CV error (confirmed by numerical analysis) when lambda is 0.03 and the polynomial degree is 3-4. Interestingly, once the degree is over 3 and there is some lambda value between 0.003 and 0.1, the error is roughly the same. Lambda seems to affect our cross-validation error more than degree, if the degree is between ~3 and 6. Of course, if the degree was extremely high with a small lambda, error would be high as the model drastically overfits the data.

Random sorting of our data into training, cross-validation, and testing sets is necessary to be sure that our model generalizes to the whole dataset, rather than fitting a certain set of examples well by chance. However, random sorting also introduces an element of unavoidable error in our data. Every time a new set of models are trained, the sorting will be different. This suggests that our hyperparameter selection may be skewed from run to run. This is exacerbated by our relatively small dataset: if we had a dataset of, say, millions of examples, any random sorting would display roughly the same general characteristics. Despite our small dataset, the same hyperparameters tended to emerge. A lambda value of roughly 0.03, and a degree of 3-4. Furthermore, the 'scoop' shape in our data tended to appear whenever we randomly sorted the data points. This is a very good sign. To confirm that our model is working as expected for the testing set, we can plot the same graph for the testing data:

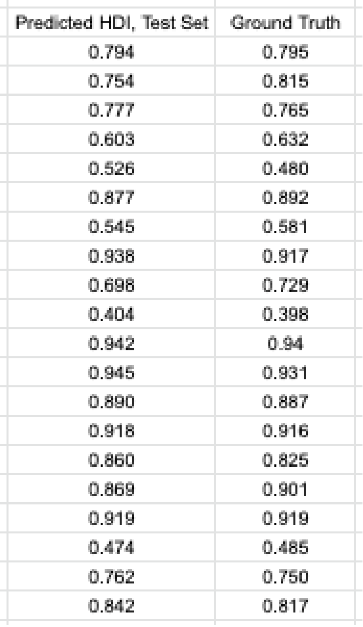


This is encouraging! We see the same shape in the training set as the testing set, and the error of the model with the best hyperparameters (=~0.0003) is approximately the same. Additionally, we know the model isn't overfitting the CV set and underfitting the test set. Finally, we can examine how the model performed on the training set, to see if it behaved as expected:



This plot is also encouraging, as it displays all of the expected errors that a training set would exhibit with variable degree and regularization. As lambda increases, we see the training error increasing as well. This is because an increase in lambda decreases the potential for overfitting, and the model does not fit the training data as closely anymore. Conversely, in the training set, an increase in degree decreases the training error because increasing the complexity of the model causes it to fit the training data very closely.

Typical predicted HDIs versus ground truth



Predicted HDI, Test Set	Ground Truth
0.794	0.795
0.754	0.815
0.777	0.765
0.603	0.632
0.526	0.480
0.877	0.892
0.545	0.581
0.938	0.917
0.698	0.729
0.404	0.398
0.942	0.94
0.945	0.931
0.890	0.887
0.918	0.916
0.860	0.825
0.869	0.901
0.919	0.919
0.474	0.485
0.762	0.750
0.842	0.817

Our final model makes accurate predictions on a country's HDI. When we examine individual examples in the test set, they are very similar (within 0.04 of the ground truth HDI). These models also indicate that the features related to women's rights correlate most to a country's HDI. This suggests that countries which have made significant feminist progress are likely to continue developing at a steady rate (given the absense of events destructive to life expectancy). Additionally, countries who true value HDI was greatly different than our predicted would benefit from increased efforts to expand women's rights. It is worth conducting additional research on the impact of individual features to determine whether the two other areas of human development we analyze–income inequality and educational enrollment–deserve to be federally prioritized above equitable freedom.

Conclusion /Future Work

The goal of our project was to determine which of our selected factors had the most influence on a nation's Human Development Index. In summary, we focused on educational attainment, women's rights, and income inequality. To construct our model, we utilized multivariate linear regression with regularization. As an improvement to our milestone model, we implemented higher-degree polynomial features to better deal with issues surrounding data complexity. We found that this implementation, alongside the optimization of our hyperparameters, produced the best results. Additionally, the rate of women's rights correlates best to a country's HDI. We suggest that those who apply our research strongly consider our dataset manually, before making an causal inference.

A next step that would be extremely useful is comparing our model to human-level benchmarks, such as by asking several experts on HDI to make predictions on a country's HDI given just the input features. Then, we could compute the same predictions using our trained and tuned regression model. Hopefully, our model could outperform the best human predictor of HDI. We would also like to further explore external variables that may have influenced our model scoring algorithm exogenously. Similarly, we could also see the effect that removing possible outliers could have on our results. Given more time and resources, it would have been fascinating to research and deploy testing programs that could have possibly diagnosed any multicollinearity within our variable scope, and to examine an emphasis on individual features in predicting the HDI.

Contributions

Milo: Machine Learning Model; Project Poster; Methods section; Experiments/Discussion Section; Abstract; Data Preprocessing Scripts; Figures and Charts; Milestone code first draft; Lambda/polynomial degree experiments; part of video presentation

Claudia: Video Slides and Presentation; portions of the final report; debugging the original milestone code/data preprocessing; researching preliminary experiments

Princess: Introduction, Related Work, Dataset, Discussion, and Conclusion sections of the final report; researching related work and significance of the project; citations; unsuccessful attempt to translate Octave code into Python

References

[1] Akin, Pelin and Tuba Koç, Tuba. 2021. "Prediction of Human Development Index with Health Indicators Using Tree-Based Regression Models." https://dergipark.org.tr/en/download/article-file/1632102.

[2] Mulyani, Sri, Andriyana, Yughie, and Sudartianto. 2017. "Modeling the human development index and the percentage of poor people using quantile smoothing splines. https://doi.org/10.1063/1.4979423.

[3] Khan, F. B. and Noor, Antika. 2021. "Prediction and Classification of Human Development Index Using Machine Learning Techniques," 5th International Conference on Electrical Information and Communication Technology (EICT), 2021, pp. 1-6, doi: 10.1109/EICT54103.2021.9733645.

[14] "Human Development Index." United Nations, Human Development Report. https://hdr.undp.org/en/composite/HDI.

[4] "Total unemployment rate (female to male)." United Nations, Human Development Report.https://hdr.undp.org/en/indicators/169706

[5] Share of seats in parliament (% held by women)." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/31706.

[6] "Gender Inequality index (GII)." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/68606.

[7] "Inequality of education (%)." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/101606.

[8] "Inequality of income (%)." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/101706.

[9] "Income share held by the richest 10%." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/187006.

[10] Expected years of schooling (years)." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/69706.

[11] Education index." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/103706.

[12] "Gross enrollment ratio, secondary (% of secondary school-age population." https://hdr.undp.org/en/indicators/63306.

[13] "Gross enrollment ratio, tertiary (% of tertiary school-age population)." United Nations, Human Development Report. https://hdr.undp.org/en/indicators/63406.