Analyzing the role of women’s rights, inequality, and education on human development around the globe

Princess Vongchanh, Claudia Kania, Milo Eirew
vongchan@stanford.edu, kania@stanford.edu, meirew@stanford.edu

Predicting

For our project, we sought to relate the human development indices of countries to global statistics involving women’s rights, inequality, and education. We built a regression model with high-order polynomial features and regularization to predict HDI given inputs on women’s rights, inequality, and education. We discovered that 4th degree polynomial features with a λ of 0.03 best represent this data. Tuning our model allowed us to train a model with an average testing error of ~0.0005. We have found definitive correlations between women’s rights, inequality, education, and HDI. Our finished model makes an extremely accurate HDI prediction, despite not being trained on the specific metrics used to calculate a country’s HDI.

Models

We are carrying out multiple linear regression with high-order polynomial features and regularization. Since our dataset is relatively small, it makes sense to use the normal equations for regularized linear regression ($\theta = [X^TX + \lambda I]^{-1} X^TY$).
In addition to the theta values that the model finds, we have two values to tune: lambda and the degree of our polynomial features. We trained a model on each value of lambda (initially up to 30, then up to 1) and each degree (up to 10) to find the lowest cross-validation error.

Discussion

Our model with higher-order polynomials, normal equations, and regularization vastly outperformed our milestone model (which used gradient descent and unregularized quadratic features). With our previous model, we obtained an average testing error of ~0.0120; now, we obtain a typical testing error of ~0.00050. This 24-fold improvement was partly due to the increased data and features we had available: we used thirty or forty (depending on the degree) features, compared to the previous six features. The fine-tuning of the regularization and polynomial degree parameters also allowed us to reduce the testing error. When we checked individual scores, the test set predictions seemed to be within around 0.03 of the real HDI. These results are very encouraging - they suggest that we have accurately related women’s rights, inequality, and education statistics to the HDI of a country using our model.

Data

We assembled our dataset using 11 different datasets from the United Nations: the columns represented the different features we used in our data, and the rows represented the countries. Each dataset had some portion of missing data, so we selected the features with minimum missing data and maximum correlation with HDI (determined by plotting each feature versus HDI). We label the ground truth for each datapoint using the known HDI of each country, and implement feature scaling for each feature.

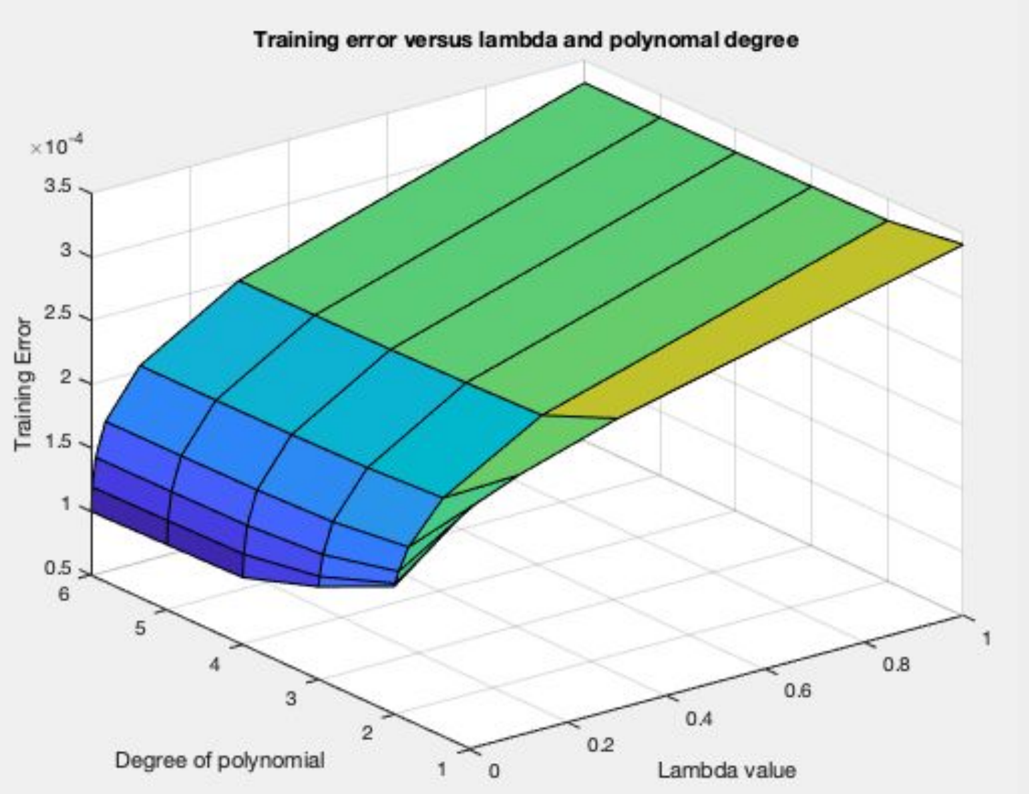
Features

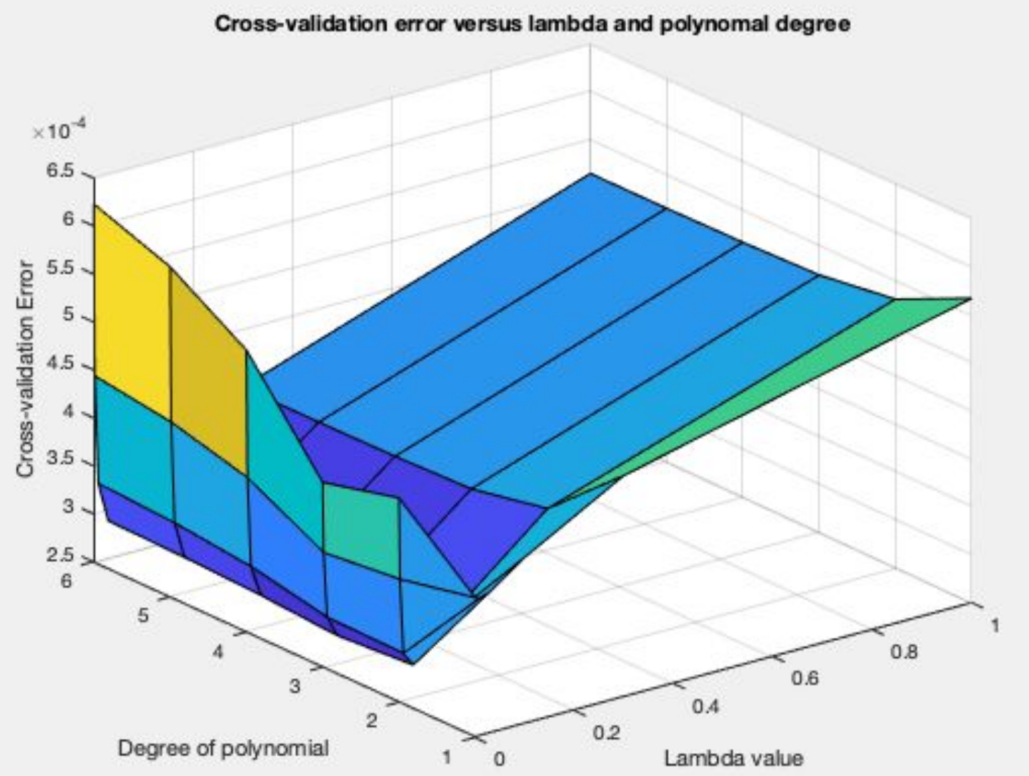
Our model is trained on thirty data features: ten from the raw input, and twenty from higher-degree polynomial features (alternately, 4th degree polynomials performed equally well). Out of the ten raw features three represent women’s rights, three represent inequality (primarily economic), and four represent education. Most features were recorded in 2019, but a small portion of the features represent an average over the last five years (a small source of error, but still notable). These features are appropriate for the task because they are representative metrics of our three topics: women’s rights, economic inequality, and education. For instance, a feature from each of the topics would be: share of seats in parliament held by women, income share held by the richest 10%, and gross enrollment ratio (secondary education).

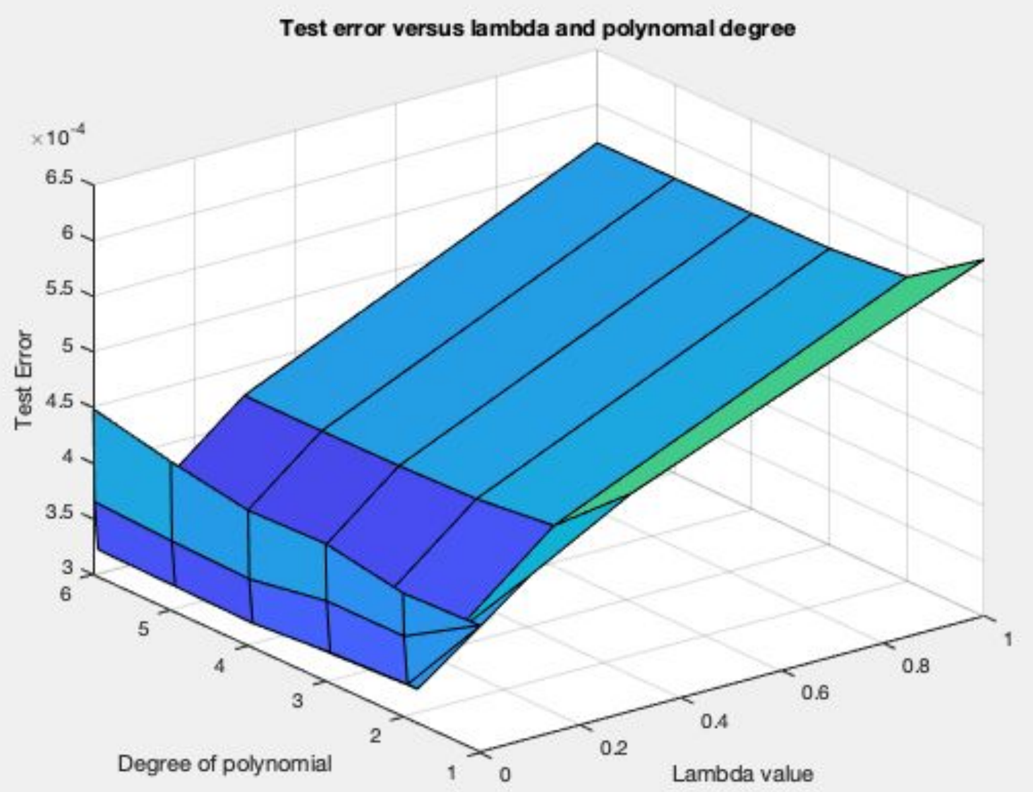
Future

I believe the limiting factor of our model is the data it is fed. Since we train the model with each datapoint representing a country, we have cap on how many datapoints exist. If we had another six weeks to develop this model, we ocus on collecting more features to use in the model, as well as refining a tool to fill in even more missing data values and seeing how other models (neural nets for instance) compare.

Plots for tuning hyperparameters







Results

The training set received 60% of the data: 60 examples. The testing set and cross-validation set (used to tune lambda and the degree of the polynomial) both received 20% of the data: 20 examples each.

Typical Test Set Predicted HDI vs Ground Truth

Predicted HDI, Test Set	Ground Truth
0.757	0.759
0.880	0.892
0.951	0.94
0.611	0.632
0.639	0.654
0.844	0.823
0.827	0.817
0.478	0.546
0.781	0.828
0.931	0.917
0.723	0.759
0.938	0.945
0.834	0.845
0.498	0.433
0.776	0.774
0.824	0.86
0.889	0.901
0.543	0.48
0.877	0.866
0.559	0.581

Model hyperparameters versus training & test error

	Training Set	Test Set
Lambda = 0.001, d = 1	0.000241	0.000414
Lambda = 0.001, d = 2	0.000151	0.000411
Lambda = 0.001, d = 3	0.000123	0.000424
Lambda = 0.001, d = 4	0.000103	0.000421
Lambda = 0.001, d = 5	0.000101	0.000433
Lambda = 0.001, d = 6	0.000100	0.000448
Lambda = 0.003, d = 1	0.000241	0.000414
Lambda = 0.003, d = 2	0.000153	0.000372
Lambda = 0.003, d = 3	0.000131	0.000372
Lambda = 0.003, d = 4	0.000121	0.000360
Lambda = 0.003, d = 5	0.000120	0.000362
Lambda = 0.003, d = 6	0.000119	0.000365
Lambda = 0.01, d = 1	0.000242	0.000416
Lambda = 0.01, d = 2	0.000161	0.000328
Lambda = 0.01, d = 3	0.000147	0.000325
Lambda = 0.01, d = 4	0.000143	0.000319
Lambda = 0.01, d = 5	0.000142	0.000320
Lambda = 0.01, d = 6	0.000141	0.000321
Lambda = 0.03, d = 1	0.000245	0.000424
Lambda = 0.03, d = 2	0.000179	0.000321
Lambda = 0.03, d = 3	0.000171	0.000318
Lambda = 0.03, d = 4	0.000169	0.000315
Lambda = 0.03, d = 5	0.000168	0.000315
Lambda = 0.03, d = 6	0.000167	0.000316
Lambda = 0.1, d = 1	0.000255	0.000447
Lambda = 0.1, d = 2	0.000210	0.000362
Lambda = 0.1, d = 3	0.000206	0.000354
Lambda = 0.1, d = 4	0.000205	0.000352
Lambda = 0.1, d = 5	0.000204	0.000351
Lambda = 0.1, d = 6	0.000204	0.000351
Lambda = 0.3, d = 1	0.000241	0.000414
Lambda = 0.3, d = 2	0.000153	0.000372
Lambda = 0.3, d = 3	0.000131	0.000372
Lambda = 0.3, d = 4	0.000121	0.000360
Lambda = 0.3, d = 5	0.000120	0.000362
Lambda = 0.3, d = 6	0.000119	0.000365
Lambda = 1, d = 1	0.000279	0.000496
Lambda = 1, d = 2	0.000254	0.000435
Lambda = 1, d = 3	0.000252	0.000427
Lambda = 1, d = 4	0.000251	0.000425
Lambda = 1, d = 5	0.000250	0.000424
Lambda = 1, d = 6	0.000250	0.000424

References

- **Related work:**
- https://github.com/julieanneco/predictingHDI
- *“Pelin AKIN , Tuba KOÇ, Prediction of Human Development Index with Health Indicators Using Tree-Based Regression Models” 2021*
- https://aip.scitation.org/doi/pdf/10.1063/1.4979423
- F. B. Khan and A. Noor, "Prediction and Classification of Human Development Index Using Machine Learning Techniques," 2021 5th International Conference on Electrical Information and Communication Technology (EICT), 2021, pp. 1-6, doi: 10.1109/EICT54103.2021.9733645.
- julieanneco, “Predicting Human Development Index using World Development Indicators and UNDP Data” 2020
- https://wallpaperaccess.com/cool-world-map

- **Data sources:**
- Gross enrollment ratio (secondary education)
- https://hdr.undp.org/en/indicators/63306
- Gross enrollment ratio (tertiary education)
- https://hdr.undp.org/en/indicators/63406
- HDI and its components:
  https://hdr.undp.org/en/composite/HDI
- Women’s empowerment - total unemployment rate (female to male) :
  https://hdr.undp.org/en/indicators/169706
- Gender Inequality index:
  https://hdr.undp.org/en/indicators/68606
- Share of seats in parliament held by women:
  https://hdr.undp.org/en/indicators/31706
- Inequality of education:
- https://hdr.undp.org/en/indicators/101606
- Inequality of income:
- https://hdr.undp.org/en/indicators/101706
- Income share held by the richest 10%
- https://hdr.undp.org/en/indicators/187006
- Education index:
- https://hdr.undp.org/en/indicators/103706
- Expected Years of Schooling
- https://hdr.undp.org/en/indicators/69706