IDinsight Works with Clients in Developing Countries to Generate and Apply Evidence for Greater Social Impact
Leveraging Existing Community and Government Resources for Girls’ Education
Policy Problem: Large # of Out of School Girls

- 3 million out of school girls (OOSGs) in India
- Caused by complex interactions between gender dynamics and economic constraints

(Source: [www.educategirls.ngo](http://www.educategirls.ngo))
Educate Girls works in 2 States: Rajasthan and Madhya Pradesh

EG planning expansion in these and other states
Problem: Where Should EG Target Program Expansion?

Current Approach:

- Select geographies based on high level admin data
- Door-to-door census for an entire district
- 2,000,000+ household surveys in 7,796 villages, in 7 districts
- Process is comprehensive, but slow and expensive

96% of Out of School Girls are in 50% of Villages in Current Program Area
Prediction Approach

Predictors:
313 Predictors
(7,796 Villages)

Outcome Variables:
Actual # of Out of School Girls
(7,796 Villages)

Results: Prediction $\rightarrow$ 50%+ efficiency gain

- Final model: random forest regression*
- We can achieve $\sim$50% increase in targeting precision over random
- A theory based OLS model yielded a $\sim$40% improvement over random
- This is an “out of sample” district that was used for final testing

*Also tested gradient descent tree models, OLS + LASSO / Ridge

Note: this is not an ROC Graph
Interesting Takeaways / Challenges / Questions?

• Takeaways for EG:
  o Can reach more out-of-school girls with less budget
  o There are additional returns to scale in investing in this work

• Takeaways for Public Policy:
  o Messy, old administrative data can still be useful for targeting
  o ML targeting opportunities are broad, not just for safety net

• Challenges & Remaining Questions:
  o Out of sample prediction complicated by geographically correlated errors. How can we better approach cross-validation?
  o How can we better aggregate predictions?
  o How should we navigate equity vs. efficiency tradeoffs when making and using predictions?
Interesting in Doing this Type of Work? We’re Hiring a Data Scientist!

Projects in the pipeline
• Improved targeting of cash transfers
• Predicting data collection quality for field surveys
• Remote sensing for infrastructure monitoring
• Predicting take-up for impact evaluations
• and more!

Posting up soon at www.idinsight.org/careers!

Questions? Contact: ben.brockman@idinsight.org
RFR Outperforms (Theory Based) OLS on Average by 5 p.p.

Rajasthan Results

Madhya Pradesh Results

5 percentage points = 5,687 additional girls identified
Predictions can identify high-burden districts

Barmer, Dungarpur, and Nagaur districts are predicted to have both a high # OOSGs/village and high total OOSGs. Programming in those districts may be highly cost-effective.