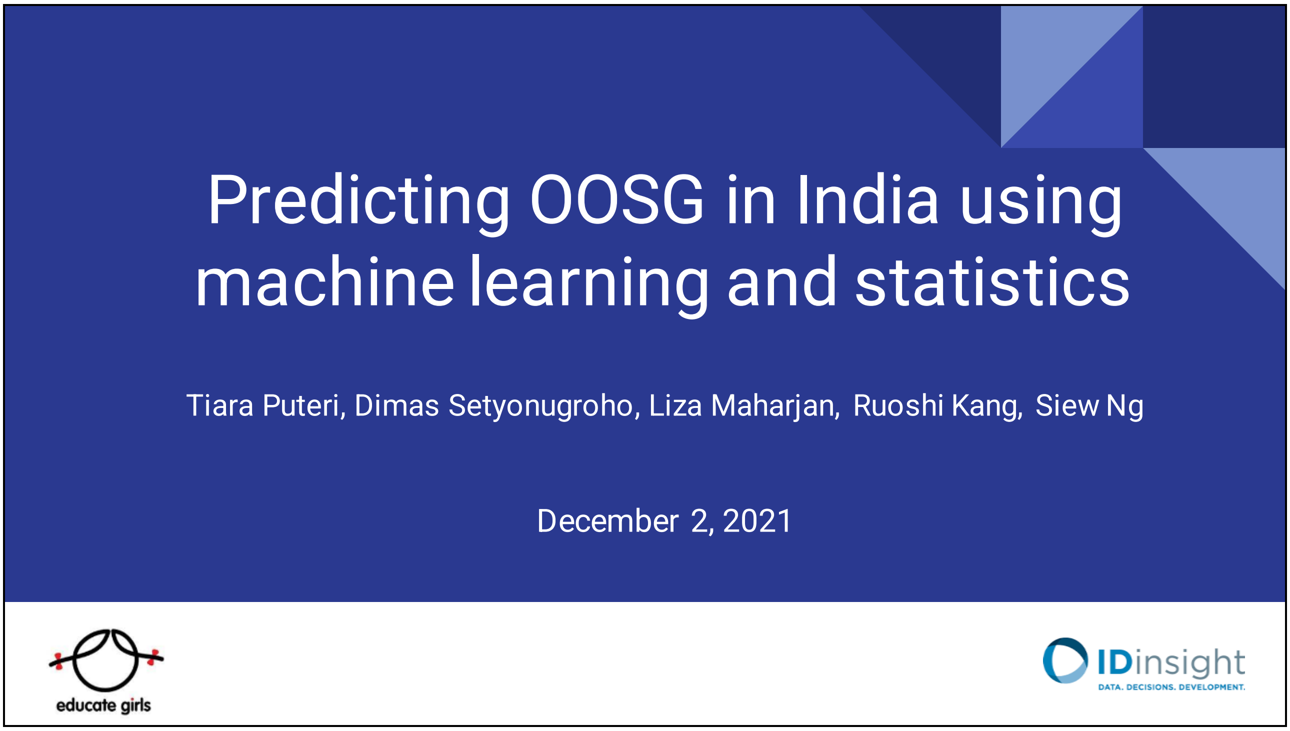


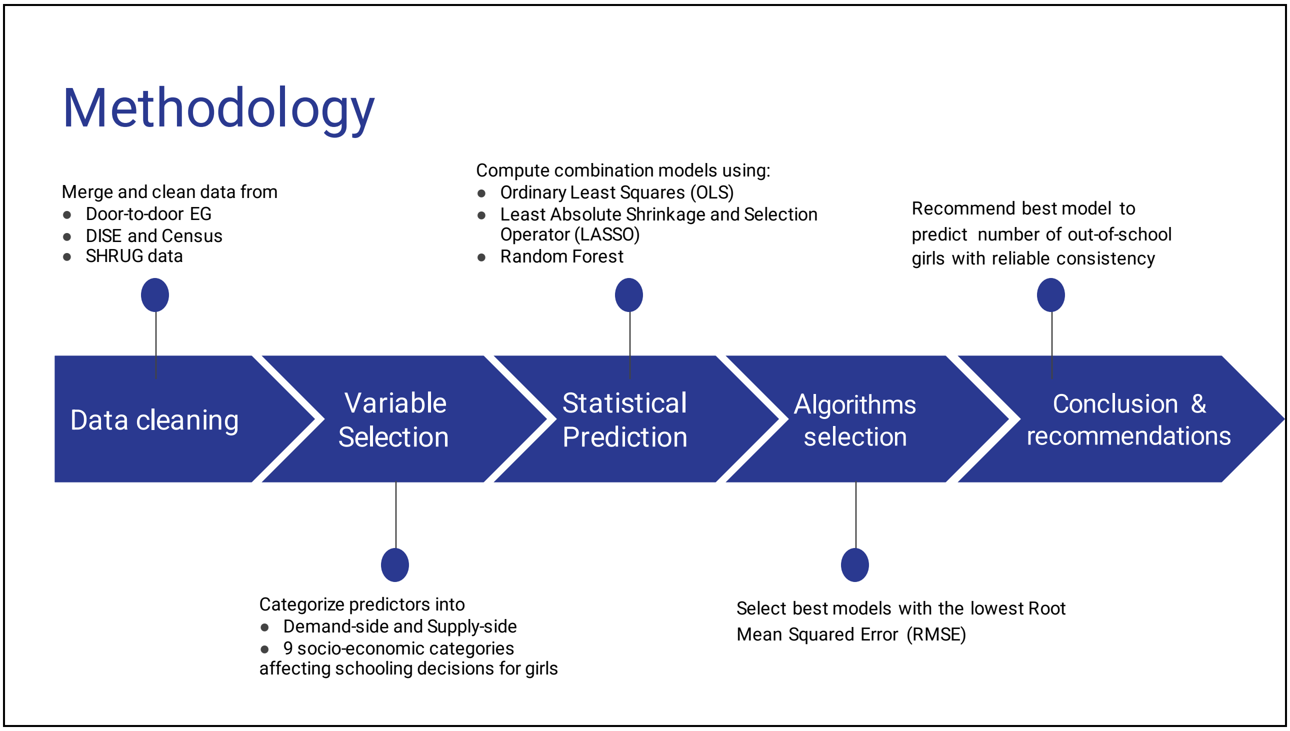
Predicting OOSG in India using machine learning and statistics

Tiara Puteri, Dimas Setyonugroho, Liza Maharjan, Ruoshi Kang, Siew Ng

December 2, 2021

Executive Summary

- This study predicts the number of out-of-school girls in each village in 14 districts in India using various statistical algorithms and machine learning techniques. The output aims to help Educate Girls (EG) improve targeting efforts in order to increase school enrollment and learning outcome for girls.
- We first evaluated 12 model candidates built using (i) Door-to-door EG data, (ii) District Information System for Education (DISE) and Village Census data, and (iii) Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) datasets with 6,538 observations and 718 variables and then, shortlisted models having the smallest Root Mean-Squared Error (RMSE).
- The recommended model is a basic **Random Forest model** with a **RMSE of 33** based on dataset consisting of only to Door-to-door EG data and DISE and Census data with 6,697 observations and 517 variables.
- However, this model does not account for inter-district variance where some districts have a significantly higher number of villages and a higher variance in number of out-of-school girls. Hence, RMSE scaled by district variance might be a better measure as compared to a simple average of RMSE across districts.



Methodology

Data cleaning

Merge and clean data from
- Door-to-door EG
- DISE and Census
- SHRUG data

Variable Selection

Categorize predictors into
- Demand-side and Supply-side
- 9 socio-economic categories affecting schooling decisions for girls

Statistical Prediction

Compute combination models using:
- Ordinary Least Squares (OLS)
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Random Forest

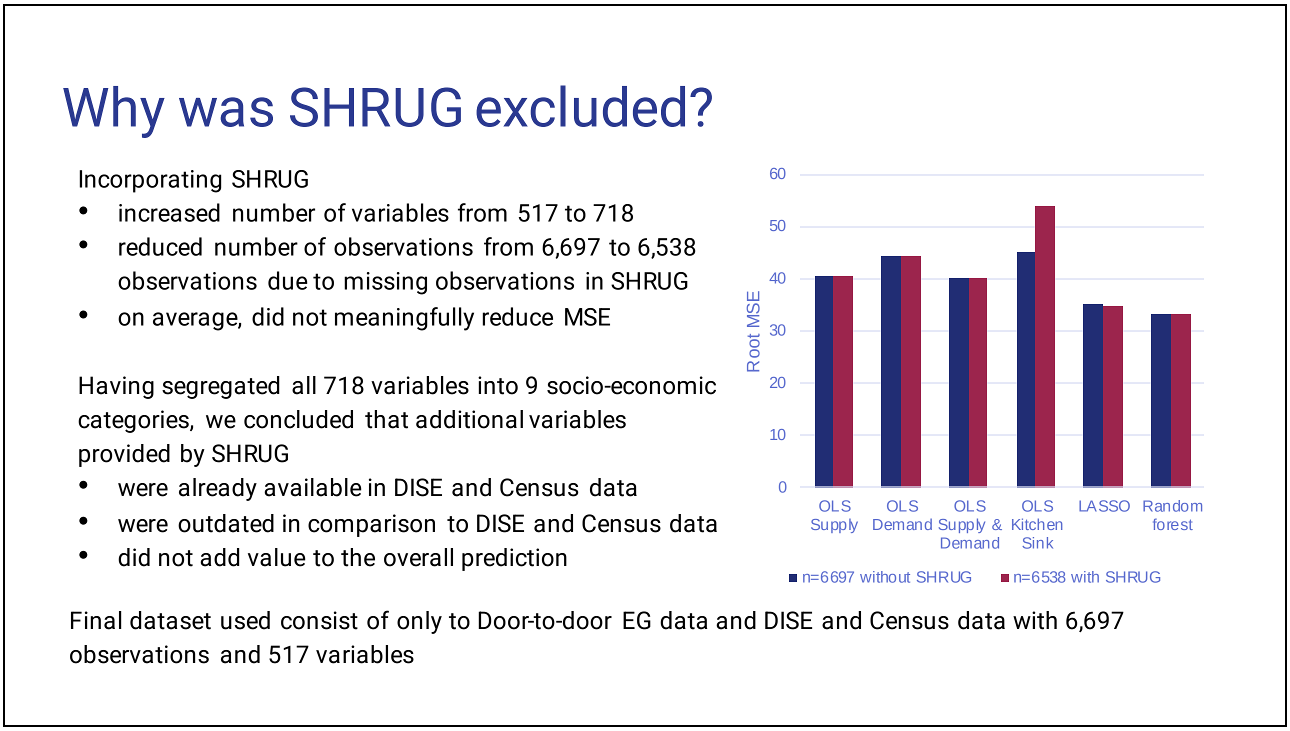
Algorithms selection

Select best models with the lowest Root Mean Squared Error (RMSE)

Conclusion & recommendations

Recommend best model to predict number of out-of-school girls with reliable consistency

In conducting the study, we went through 5 steps, starting with data cleaning, selecting the variables, followed by conducting statistical prediction using models such as OLS, LASSO, and Random Forest. After that, we select the best models with the lowest Root Mean Squared Error (RMSE). After selecting the best algorithm, we derive conclusion and recommendations on which villages Education Girls should prioritize in the program.



Why was SHRUG excluded?

Incorporating SHRUG

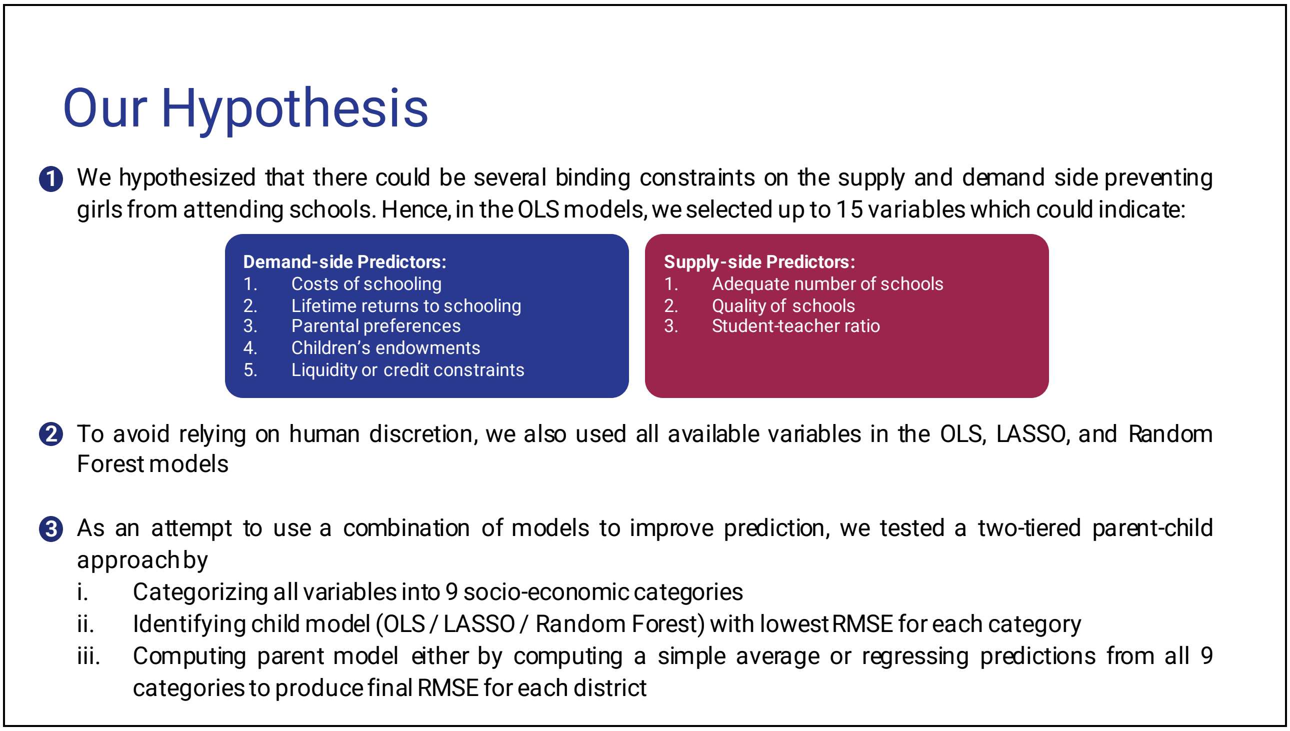
- increased number of variables from 517 to 718
- reduced number of observations from 6,697 to 6,538 observations due to missing observations in SHRUG
- on average, did not meaningfully reduce MSE

Having segregated all 718 variables into 9 socio-economic categories, we concluded that additional variables provided by SHRUG

- were already available in DISE and Census data
- were outdated in comparison to DISE and Census data
- did not add value to the overall prediction

Final dataset used consist of only to Door-to-door EG data and DISE and Census data with 6,697 observations and 517 variables

In going through the data cleaning and processing, we tried merging census, door-to-door EG, DISE, and SHRUG data. However we found that the additional variables found in SHRUG were already in the DISE and Census data, thus this would be redundant. Moreover, the additional data from SHRUG resulted in less observations but higher RMSE, did not add significant value to the overall production, thus we decided not to use SHRUG and only use Door-to-door EG and DISE and census data



Our Hypothesis

❶ We hypothesized that there could be several binding constraints on the supply and demand side preventing girls from attending schools. Hence, in the OLS models, we selected up to 15 variables which could indicate:

Demand-side Predictors:

1. Costs of schooling
2. Lifetime returns to schooling
3. Parental preferences
4. Children's endowments
5. Liquidity or credit constraints

Supply-side Predictors:

1. Adequate number of schools
2. Quality of schools
3. Student-teacher ratio

❷ To avoid relying on human discretion, we also used all available variables in the OLS, LASSO, and Random Forest models

❸ As an attempt to use a combination of models to improve prediction, we tested a two-tiered parent-child approach by

i. Categorizing all variables into 9 socio-economic categories

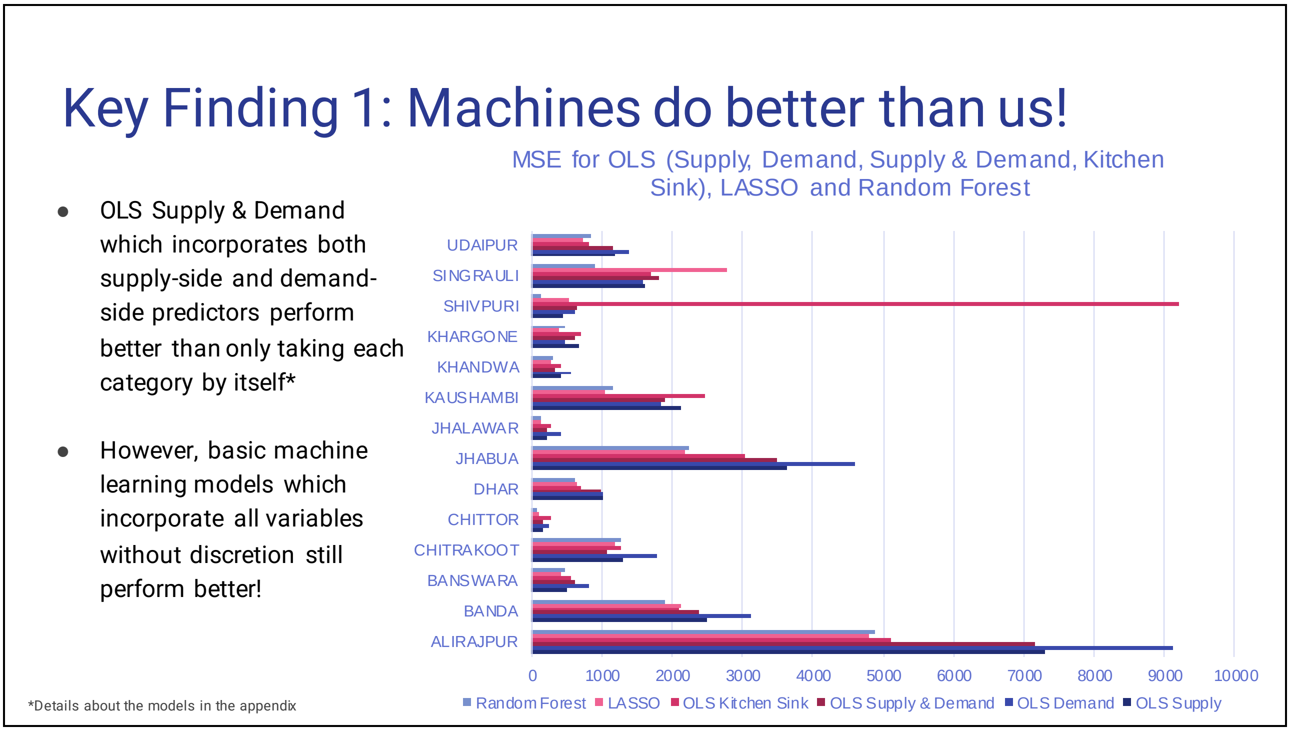
ii. Identifying child model (OLS / LASSO / Random Forest) with lowest RMSE for each category

iii. Computing parent model either by computing a simple average or regressing predictions from all 9 categories to produce final RMSE for each district

We hypothesize there could be varioindicate demand side constraints such as: costs of schooling (e.g., average school fees, opportunity cost of time like foregone employment), lifetime returns to schooling (e.g., earnings returns, information about returns, access to profitable employment opportunities, life expectancy), parental preferences (e.g., literacy rate of parents, son vs daughter, preference based on other observable and non-observable characteristics), children's endowments (e.g., IQ), liquidity/credit constraints (family assets, earnings, access to financial institutions etc.)

And similarly up to 10 supply side constraints: school approachability by road, school evaluation, electricity in school, English as school's medium of instruction, availability of mid-day meals in school, number of schools in village, student teacher ratio, number of girls toilet in school, textbook availability in school, and availability of water in school.

To avoid relying on human discretion, we used all available variables in the OLS, LASSO, and Random Forest models.
We also used two tier parent child models leveraging 9 categories of variables and various ways such as simple averaging and regressing predictions from the nine categories to produce a final prediction for each villages.



Key Finding 1: Machines do better than us!

- OLS Supply & Demand which incorporates both supply-side and demand-side predictors perform better than only taking each category by itself*
- However, basic machine learning models which incorporate all variables without discretion still perform better!

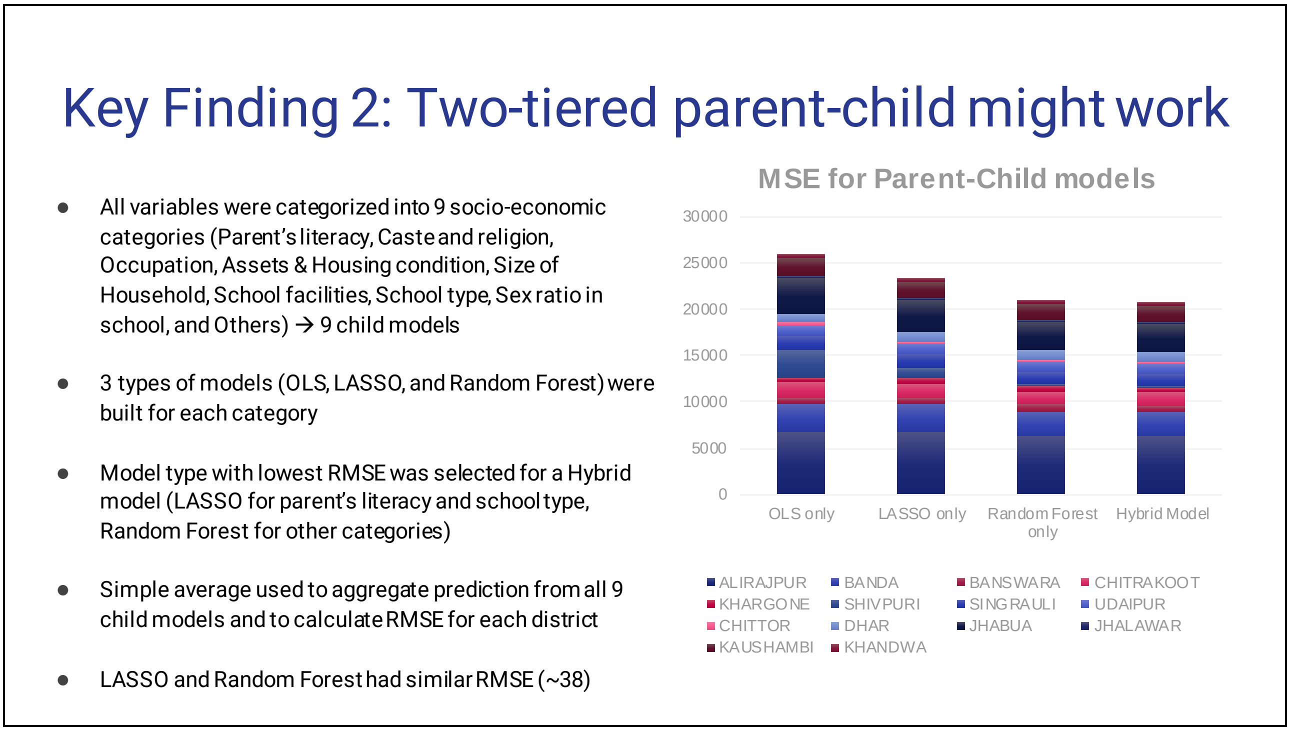
*Details about the models in the appendix

Based on the hypothesised demand-side predictors, we shortlisted 12 variables indicating demand-side constraints. Based on the hypothesised supply-side predictors, we shortlisted 10 variables indicating supply-side constraints.

An OLS model was developed each for the (1) demand-side predictors, (2) supply-side predictors, and (3) a combination of demand and supply side predictors. The model combining demand and supply side predictors produced the lowest average MSE of 1,604.9 for the holdout sample compared to only using demand-side or supply-side predictors.

However, since the variables were shortlisted based on discretion, significant variables may be omitted. Using All Available Variables To mitigate selection bias, all available variables were used to develop an OLS model (termed as ‘kitchen sink’), a LASSO model and a Random Forest model.

Random Forest using all variables produced the lowest average MSE of 1,092.6 and we find that machine learning techniques yield better predictions than OLS models.



Key Finding 2: Two-tiered parent-child might work

- All variables were categorized into 9 socio-economic categories (Parent's literacy, Caste and religion, Occupation, Assets & Housing condition, Size of Household, School facilities, School type, Sex ratio in school, and Others) → 9 child models
- 3 types of models (OLS, LASSO, and Random Forest) were built for each category
- Model type with lowest RMSE was selected for a Hybrid model (LASSO for parent's literacy and school type, Random Forest for other categories)
- Simple average used to aggregate prediction from all 9 child models and to calculate RMSE for each district
- LASSO and Random Forest had similar RMSE (~38)

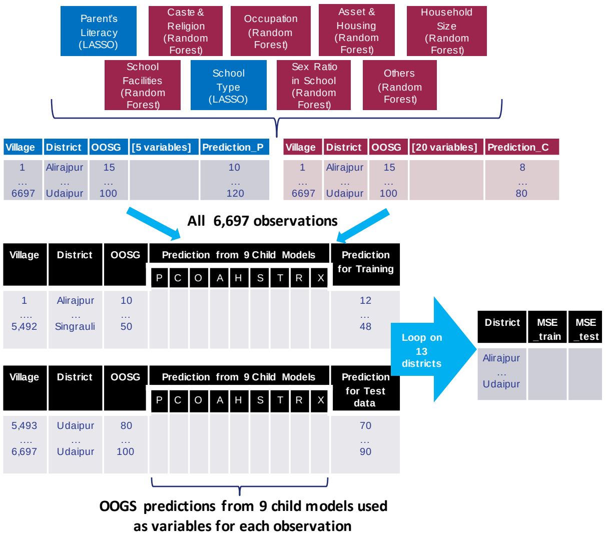
Further, we explored categorizing all variables into 9 socio-economic categories and fitting models on all the categories separately and combining them to yield better predictions. In this two-tiered Parent-Child Approach, the 9 socio-economic categories are: (1) parent's literacy, (2) caste and religion, (3) occupation, (4) assets and housing condition, (5) household size, (6) school facilities, (7) school type, (8) sex ratio in school, and (9) other factors. This creates 9 individual child models.

To aggregate the predictions from all 9 child models, we explored 4 different methods for the parent model.

Method 1: For each category, all three types of models (OLS, LASSO, and Random Forest) were developed. The MSE for each district was computed based on a simple average according to each model type. Of all model types, Random Forest yields the lowest MSE for the test data.

Method 2: For each category, the model type which yields the lowest MSE was selected. For category (1) parent's literacy and (7) school type, LASSO was selected and for all other categories, Random Forest was selected. The MSE for each district was computed based on a simple average, which is lower than only using Random Forest as the model type in Method 1.

Key Finding 3: Predictions on top of prediction!



Parent's Literacy (LASSO)

Caste & Religion (Random Forest)

Occupation (Random Forest)

Asset & Housing (Random Forest)

Household Size (Random Forest)

School Facilities (Random Forest)

School Type (LASSO)

Sex Ratio in School (Random Forest)

Others (Random Forest)

Village	District	OOSG	[5 variables]	Prediction_P
1 … 6697	Alirajpur … Udaipur	15 … 100		10 … 120

Village	District	OOSG	[20 variables]	Prediction_C
1 … 6697	Alirajpur … Udaipur	15 … 100		8 … 80

All 6,697 observations

Village	District	OOSG	Prediction from 9 Child Models									Prediction for Training
			P	C	O	A	H	S	T	R	X	
1 …. 5,492	Alirajpur … Singrauli	10 … 50										12 … 48

Village	District	OOSG	Prediction from 9 Child Models									Prediction for Test data
			P	C	O	A	H	S	T	R	X	
5,493 …. 6,697	Udaipur … Udaipur	80 … 100										70 … 90

Loop on 13 districts

District	MSE _train	MSE _test
Alirajpur … Udaipur		

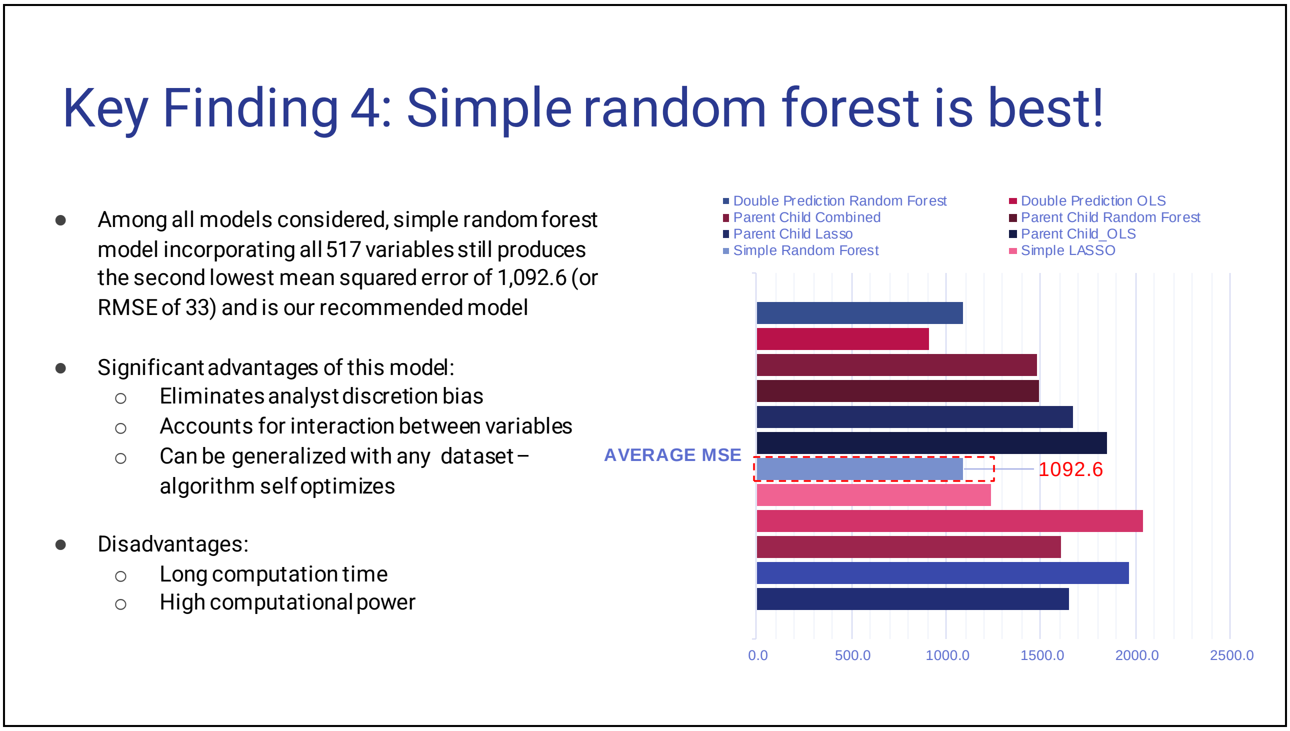
OOGS predictions from 9 child models used as variables for each observation

Hybrid of LASSO and Random Forest, using OLS for weighted average

- Simple average might not be representative of importance of different child models, but giving weightage might be arbitrary
- To eliminate manual assignment of weights for each child model, we regressed OOSG predictions from the 9 child models to actual OOSG
- This method yields the lowest MSE of 913.5 for the test data, compared to all other methods
- But hierarchical form of modelling introduces more uncertainty and complicates construction of confidence interval

To aggregate the predictions from all 9 child models, we explored 4 different methods for the parent model. [method discussed above]

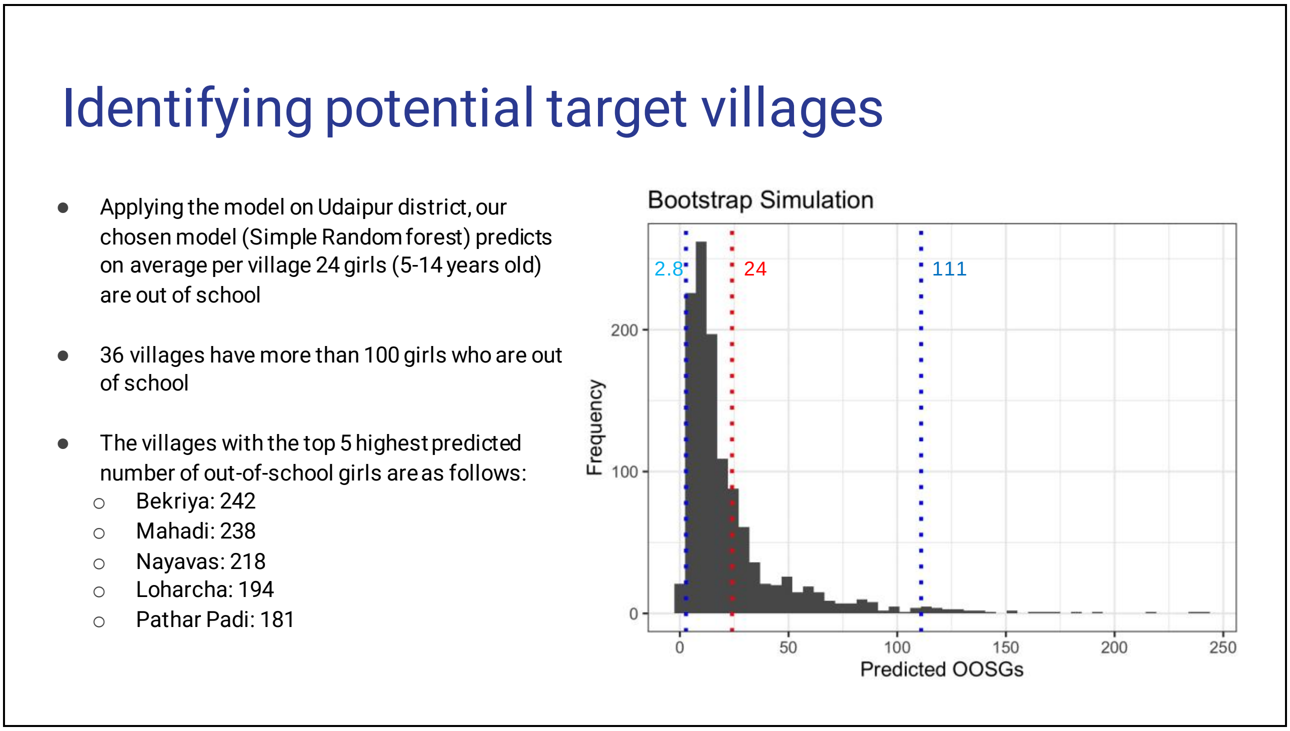
Method 3: Combination of LASSO and Random Forest models as per Method 2 were developed, using all available observations. Instead of using a simple average, an OLS model is used to regress the predictions and calculate the weightage for each child model. This method yields the lowest MSE of 913.5 for the test data, compared to all other methods.
Method 4: Combination of LASSO and Random Forest models as per Method 2 were developed, using only observations from 13 districts. Instead of using a simple average, a Random Forest model is used to regress the predictions and calculate the weightage for each child model.



Key Finding 4: Simple random forest is best!

- Among all models considered, simple random forest model incorporating all 517 variables still produces the second lowest mean squared error of 1,092.6 (or RMSE of 33) and is our recommended model
- Significant advantages of this model:
  - Eliminates analyst discretion bias
  - Accounts for interaction between variables
  - Can be generalized with any dataset – algorithm self optimizes
- Disadvantages:
  - Long computation time
  - High computational power

- Among all models considered, simple random forest model incorporating all 517 variables still produces the second lowest mean squared error of 1,092.6 (or RMSE of 33). The only model that performs better than simple random forest and has the lowest MSE is the two-tiered parent-child OLS double prediction model with an MSE of 914.
- However, we still recommend the simple Random forest model because of its many advantages: One, it eliminates analyst discretion bias hence it can be easily generalized with any dataset and the algorithm self optimizes to best fit the data The model also accounts for interaction between variables hence However, it does have a few downsides including long computation time and significant computational power required



Identifying potential target villages

- Applying the model on Udaipur district, our chosen model (Simple Random forest) predicts on average per village 24 girls (5-14 years old) are out of school
- 36 villages have more than 100 girls who are out of school
- The villages with the top 5 highest predicted number of out-of-school girls are as follows:
  - ○ Bekriya: 242
  - ○ Mahadi: 238
  - ○ Nayavas: 218
  - ○ Loharcha: 194
  - ○ Pathar Padi: 181

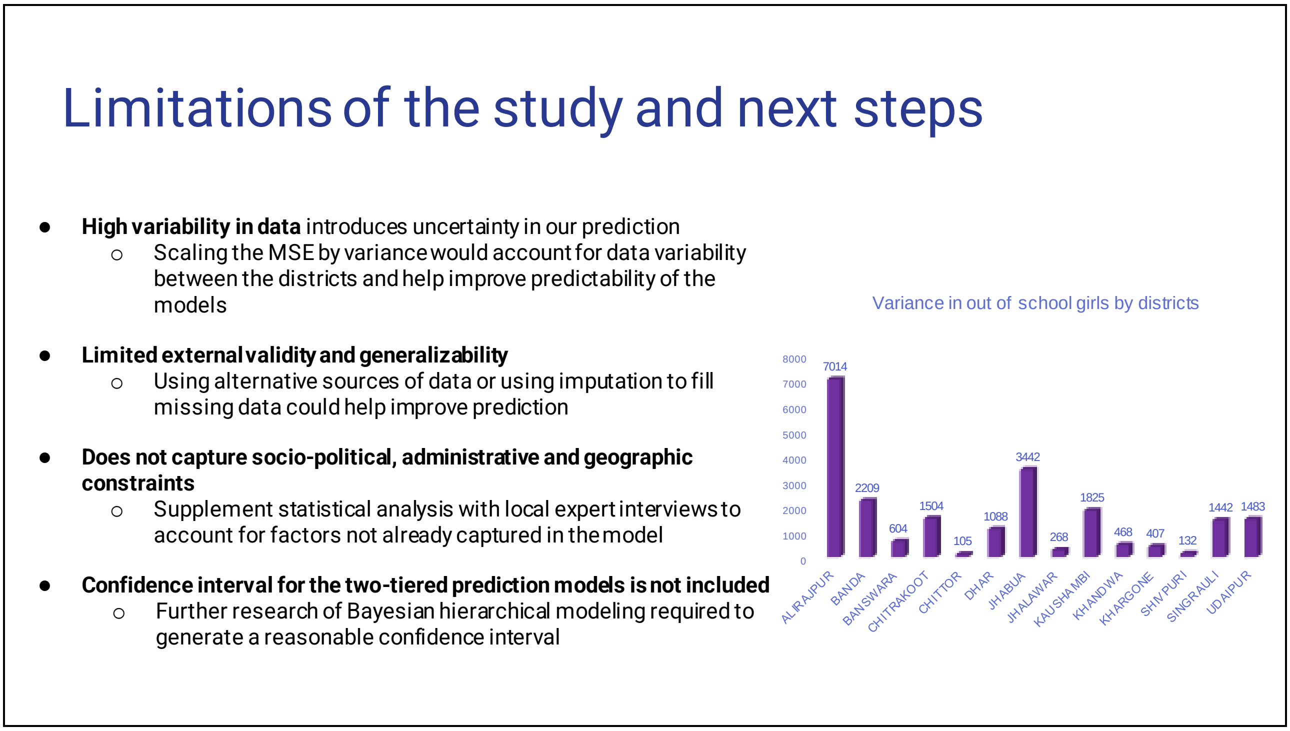
Using our chosen simple random forest model, we predicted out of school girls in Udaipur district. This graph on the right shows the predicted OOGS for the whole of Udaipur district with an average of 24, and lower and upper bound of 3 and 111 girls respectively. This wide confidence interval can be explained by the right skew in our predicted data.

A few things to note:

1. We have only ~1199 villages in Udaipur we're predicting on. Some of the villages with missing values were eliminated during data cleaning process
2. We were able to bootstrap only 100 times (due to time and computational power constraints)

A few other things to highlight: Our model predicts 36 villages have more than 100 school going girls who are out of school. Out of these villages, the villages with the top 5 highest predicted number of out-of-school girls are as follows:

- ○ Bekriya: 242
- ○ Mahadi: 238
- ○ Nayavas: 218
- ○ Loharcha: 194
- ○ Pathar Padi: 181



Limitations of the study and next steps

- **High variability in data** introduces uncertainty in our prediction
  - Scaling the MSE by variance would account for data variability between the districts and help improve predictability of the models
- **Limited external validity and generalizability**
  - Using alternative sources of data or using imputation to fill missing data could help improve prediction
- **Does not capture socio-political, administrative and geographic constraints**
  - Supplement statistical analysis with local expert interviews to account for factors not already captured in the model
- **Confidence interval for the two-tiered prediction models is not included**
  - Further research of Bayesian hierarchical modeling required to generate a reasonable confidence interval

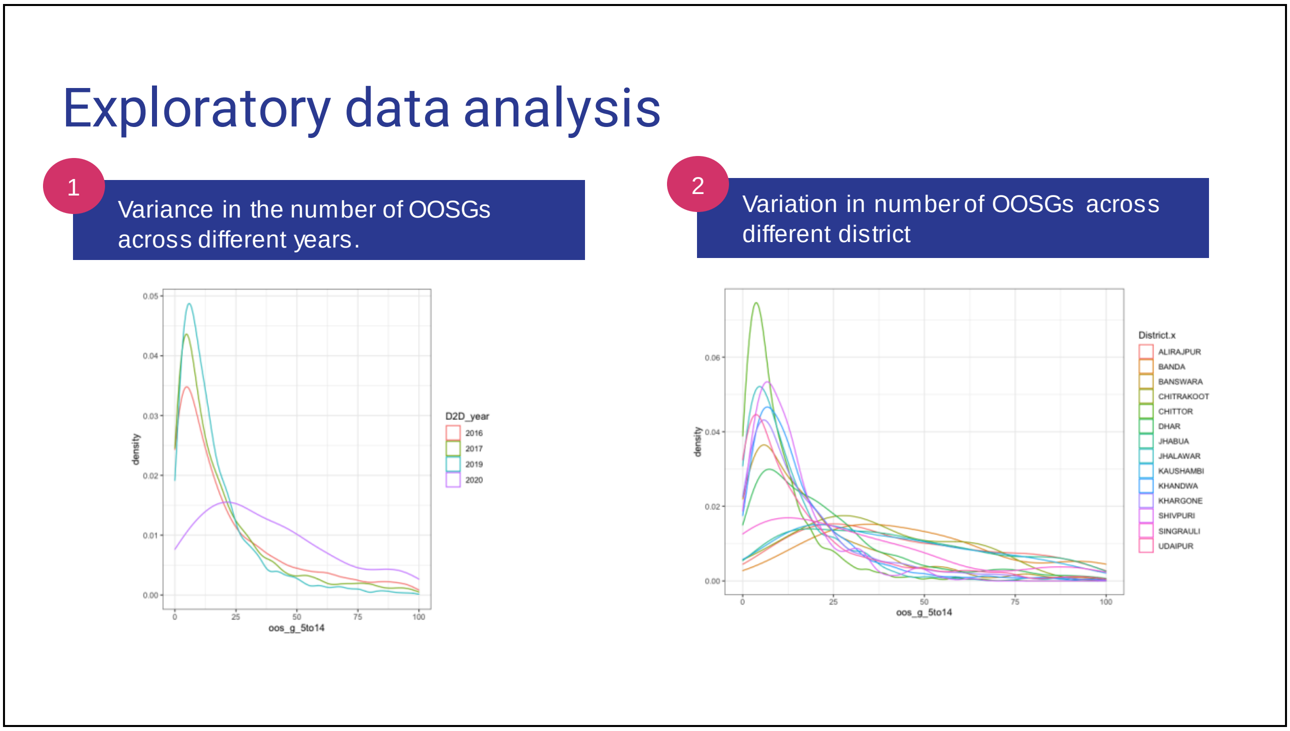
Variance in out of school girls by districts

- High variability in data (number of girls out of school and number of villages in each district) introduces uncertainty in our prediction. Scaling the MSE by variance would account for data variability between the districts and help improve predictability of the models
- Limited external validity and generalizability since our model is based on data from 14 out of 748 districts (<2%) and the set of variables used in the models are not comprehensive. Having a more complete data or using imputation to retain could help improve prediction.
- Does not capture socio-political, administrative and geographic constraints not already included in data. Supplement statistical analysis with local expert interviews to account for factors not already captured
- Confidence interval for the two-tiered prediction models is not included - two levels of uncertainty in the two-tiered parent-child approach. This requires further research of Bayesian hierarchical modeling to generate a reasonable confidence interval

Technical Appendix

Data

Dataset	Description
Door-to-door (D2D) data from Educate Girls (EG)	This data has numbers of out-of-school girls aged 5 to 14 from 8,980 villages,14 district in India from 2016, 2017, 2019, and 2020.
District Information System for Education (DISE)	This data covers education-related indicators across 21,044 villages in India. It also has different sub data: Basic information, General information, enrollment data, Facility data, RTE data, teaching data.
Census data	Census data covers extensive information on socio-economic characteristics of the village. In general, the census data has: Assets ownership among households in each village, Cooking fuel used by households in each village, etc.
Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) data	SHRUG data contains extensive socioeconomic data across the whole regions of India, such as Socio-economic caste census data, Electrification, Non-farm employment, Road and remote sensing, and Night lights data



Exploratory data analysis

1 Variance in the number of OOSGs across different years.

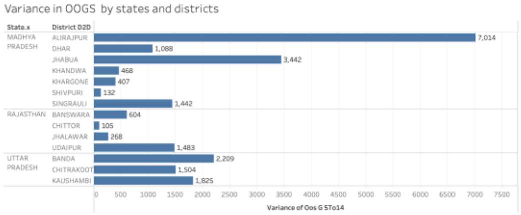
2 Variation in number of OOSGs across different district

In the 6697 observations, there is no substantial variation in OOSGs across different years, especially across 2016, 2017, and 2019. We can neglect the differences that are visually visible in 2020 since the data points in this year are very few (i.e. 162) relative to 2016 (i.e. 2,170), 2017 (i.e. 3,075), and 2019 (i.e. 1,290). There is a variation in OOSGs across different districts. Some districts like Alirajpur, Banda, Chitrakoot, Jhabua, Kaushambi, and Singrauli, on average, seem to have a more intensive number of OOSGs relative to other districts.

Exploratory data analysis (continued)

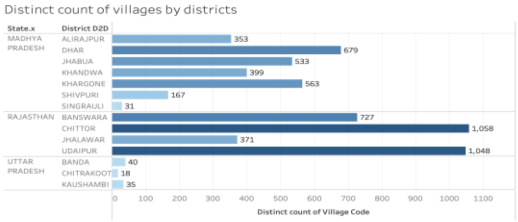
3

Variance in number of OOSGs across different districts

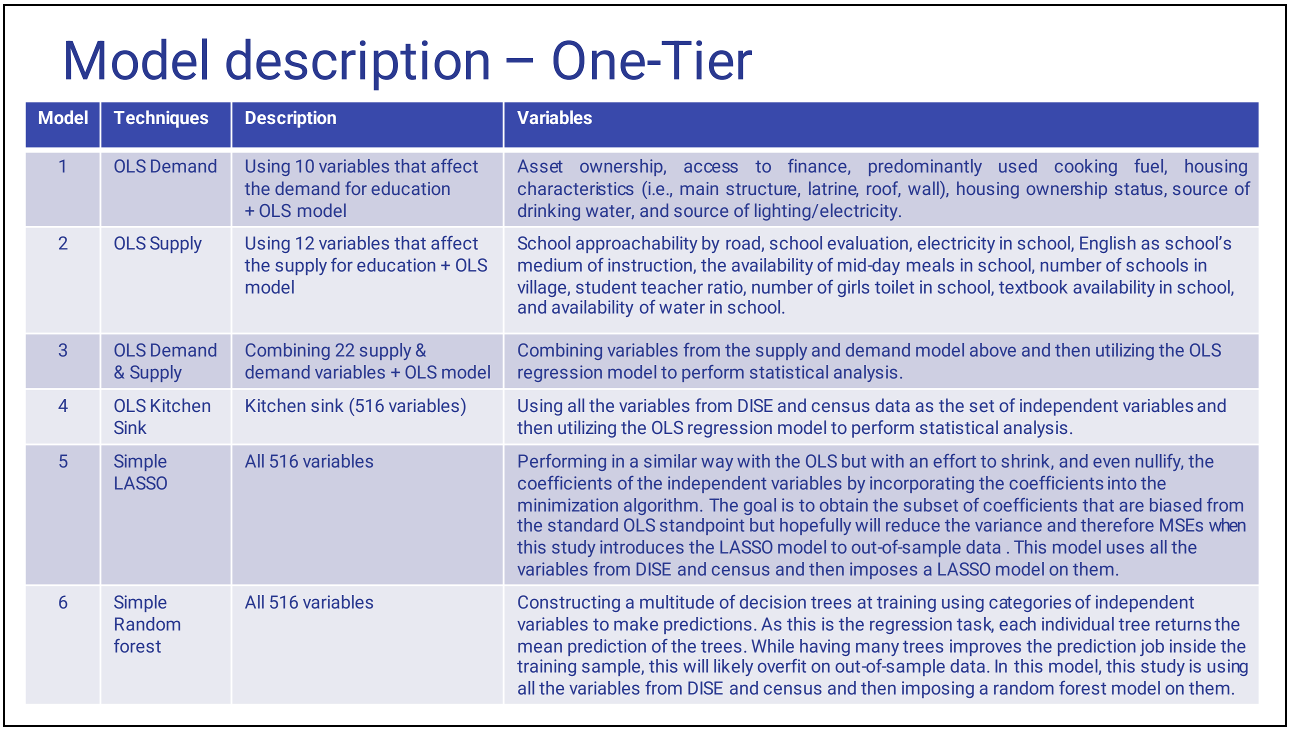


4

Number of villages per district



The highest variance is found in the Madhya Pradesh state, especially in the Alirajpur district, with variance of 7,014; followed by Jhabua, and Bandi. Regarding the number of villages, the observations mostly come from Madhya Pradesh and Rajasthan states with the most number of villages.



Model description – One-Tier

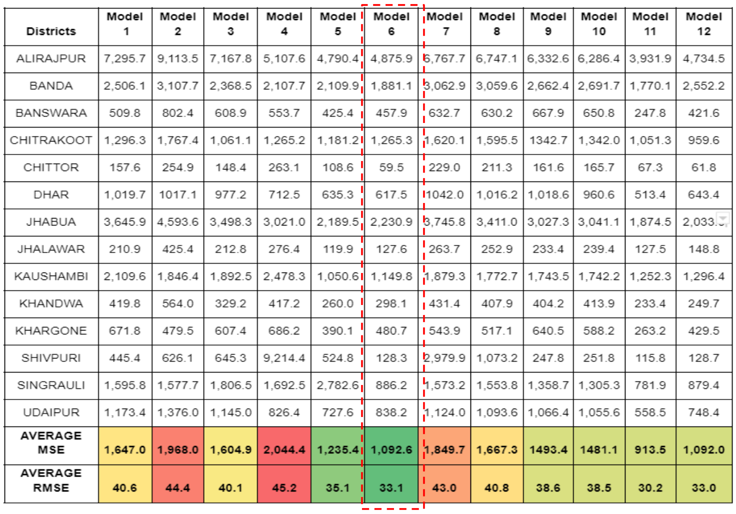
Model	Techniques	Description	Variables
1	OLS Demand	Using 10 variables that affect the demand for education + OLS model	Asset ownership, access to finance, predominantly used cooking fuel, housing characteristics (i.e., main structure, latrine, roof, wall), housing ownership status, source of drinking water, and source of lighting/electricity.
2	OLS Supply	Using 12 variables that affect the supply for education + OLS model	School approachability by road, school evaluation, electricity in school, English as school's medium of instruction, the availability of mid-day meals in school, number of schools in village, student teacher ratio, number of girls toilet in school, textbook availability in school, and availability of water in school.
3	OLS Demand & Supply	Combining 22 supply & demand variables + OLS model	Combining variables from the supply and demand model above and then utilizing the OLS regression model to perform statistical analysis.
4	OLS Kitchen Sink	Kitchen sink (516 variables)	Using all the variables from DISE and census data as the set of independent variables and then utilizing the OLS regression model to perform statistical analysis.
5	Simple LASSO	All 516 variables	Performing in a similar way with the OLS but with an effort to shrink, and even nullify, the coefficients of the independent variables by incorporating the coefficients into the minimization algorithm. The goal is to obtain the subset of coefficients that are biased from the standard OLS standpoint but hopefully will reduce the variance and therefore MSEs when this study introduces the LASSO model to out-of-sample data . This model uses all the variables from DISE and census and then imposes a LASSO model on them.
6	Simple Random forest	All 516 variables	Constructing a multitude of decision trees at training using categories of independent variables to make predictions. As this is the regression task, each individual tree returns the mean prediction of the trees. While having many trees improves the prediction job inside the training sample, this will likely overfit on out-of-sample data. In this model, this study is using all the variables from DISE and census and then imposing a random forest model on them.

Model description – Two-tiered Parent-Child

Model	Techniques	Variables	Description
7	OLS	Grouping all 516 variables into 9 different groups: 1. Parent's literacy 2. Caste and religion, Occupation 3. Assets and housing condition 4. Size of Household 5. School facilities 6. School type 7. Sex ratio in school 8. Other factors	For each category, an OLS model was developed. The MSE for each district was computed based on a simple average of all 9 child models.
8	LASSO		For each category, a LASSO model was developed. The MSE for each district was computed based on a simple average of all 9 child models.
9	Random Forest		For each category, a Random Forest model was developed. The MSE for each district was computed based on a simple average of all 9 child models.
10	Hybrid		For category (1) parent's literacy and (7) school type, LASSO was used and for all others, Random Forest was used. The MSE for each district was computed based on a simple average.
11	Hybrid + OLS for weighted average		Combination of LASSO and Random Forest models as per Model 10 were developed, using all available observations. Instead of using a simple average, an OLS model is used to regress the predictions and calculate the weightage for each child model. Only at the parent model stage, the OLS model was developed using a training dataset of 13 districts and tested on the remaining district.
12	Hybrid + LASSO for weighted average		Combination of LASSO and Random Forest models as per Model 10 were developed. Instead of using a simple average, a Random Forest model is used to regress the predictions and calculate the weightage for each child model.

* Except for Model 11, all models were developed using a training dataset of 13 districts and tested on the remaining district. The loop is repeated for every district.

MSEs across models



Districts	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
ALIRAJPUR	7,295.7	9,113.5	7,167.8	5,107.6	4,790.4	4,875.9	6,767.7	6,747.1	6,332.6	6,286.4	3,931.9	4,734.5
BANDA	2,506.1	3,107.7	2,368.5	2,107.7	2,109.9	1,881.1	3,062.9	3,059.6	2,662.4	2,691.7	1,770.1	2,552.2
BANSWARA	509.8	802.4	608.9	553.7	425.4	457.9	632.7	630.2	667.9	650.8	247.8	421.6
CHITRAKOOT	1,296.3	1,767.4	1,061.1	1,265.2	1,181.2	1,265.3	1,620.1	1,595.5	1342.7	1,342.0	1,051.3	959.6
CHITTOR	157.6	254.9	148.4	263.1	108.6	59.5	229.0	211.3	161.6	165.7	67.3	61.8
DHAR	1,019.7	1017.1	977.2	712.5	635.3	617.5	1042.0	1,016.2	1,018.6	960.6	513.4	643.4
JHABUA	3,645.9	4,593.6	3,498.3	3,021.0	2,189.5	2,230.9	3,745.8	3,411.0	3,027.3	3,041.1	1,874.5	2,033.[illegible]
JHALAWAR	210.9	425.4	212.8	276.4	119.9	127.6	263.7	252.9	233.4	239.4	127.5	148.8
KAUSHAMBI	2,109.6	1,846.4	1,892.5	2,478.3	1,050.6	1,149.8	1,879.3	1,772.7	1,743.5	1,742.2	1,252.3	1,296.4
KHANDWA	419.8	564.0	329.2	417.2	260.0	298.1	431.4	407.9	404.2	413.9	233.4	249.7
KHARGONE	671.8	479.5	607.4	686.2	390.1	480.7	543.9	517.1	640.5	588.2	263.2	429.5
SHIVPURI	445.4	626.1	645.3	9,214.4	524.8	128.3	2,979.9	1,073.2	247.8	251.8	115.8	128.7
SINGRAULI	1,595.8	1,577.7	1,806.5	1,692.5	2,782.6	886.2	1,573.2	1,553.8	1,358.7	1,305.3	781.9	879.4
UDAIPUR	1,173.4	1,376.0	1,145.0	826.4	727.6	838.2	1,124.0	1,093.6	1,066.4	1,055.6	558.5	748.4
AVERAGE MSE	**1,647.0**	**1,968.0**	**1,604.9**	**2,044.4**	**1,235.4**	**1,092.6**	**1,849.7**	**1,667.3**	**1493.4**	**1481.1**	**913.5**	**1,092.0**
AVERAGE RMSE	**40.6**	**44.4**	**40.1**	**45.2**	**35.1**	**33.1**	**43.0**	**40.8**	**38.6**	**38.5**	**30.2**	**33.0**

- Model 6, 11, and 12 have the lowest RMSEs
- Model 6, Simple Random Forest, is selected as the recommended model for Educate Girls