STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's Thesis in Economics
Academic Year 2018-19

Extraordinary or Ordinary at Best?

An Empirical Study on the Application of Machine Learning Tools for Proxy Means Tests in Poverty Targeting

Nicolas Leicht (41207) and Colja Maser (41224)

Abstract

Proxy means tests are a widely used approach in development programs where the beneficiaries need to be determined through targeting. These tests apply a standard econometric method, ordinary least squares, to predict consumption levels using household characteristics as input variables. Yet, they still exhibit substantial misclassification rates when it comes to determining whether a household is poor or not. In this thesis, we investigate whether three alternative statistical approaches, penalized regressions, random forests or neural networks, could be applied to decrease these misclassification rates. For this purpose, we use two multi-topic household panel surveys from India and Indonesia and apply an out-of-sample validation procedure. Additionally, we evaluate how good the different methods predict poverty over time. While neural networks yield the lowest misclassification rates for most of our analyses, overall, we conclude that the precision of the methods does not differ from each other both from a statistical and economic perspective. These results are robust for important subgroups, a different set of input variables and a lower poverty line. Additionally, we find that the targeting accuracy of all methods is very stable over time.

Keywords: Poverty targeting, proxy means tests, machine learning, household surveys

JEL: C14, C45, I32, I38, O15

Supervisor: Abhijeet Singh
Date submitted: 13.05.2019
Date examined: 27.05.2019
Discussant: Sean Tay
Examiner: Mark Sanctuary

Acknowledgements

First of all, we would like to thank our supervisor Abhijeet Singh for his guidance throughout this term. In particular, we valued his feedback on developing our research question and his challenging opinions, which have significantly contributed to our work.

Additionally, we thank Rickard Sandberg who was willing to advise us on the more technical aspects of our thesis. His comments were reassuring and greatly appreciated.

And finally, both of us are more than grateful to our families who have given us the opportunity to follow our interests and given us comfort during all times.

Contents

List of Tables

List of Figures

List of Abbreviations

BLT	Bantuan Langsung Tunai
FE	Fixed Effects
GDP	Gross Domestic Product
IFLS	Indonesian Family Life Survey
IHDS	India Human Development Survey
NN	Neural Network
OLS	Ordinary Least Squares
PMT	Proxy Means Test
PR	Penalized Regression
RF	Random Forest
USAID	United States Agency for International Development

1 Introduction

Many developing countries employ large scale cash transfer programs to fight poverty. Prominent programs are *Oportunidades* in Mexico, *Familias en Acción* in Colombia or the *Bantuan Langsung Tunai (BLT)* program in Indonesia. While these programs vary in exact set-up and size, they all use targeting mechanisms and thus only aim at a certain segment of the population. Targeting has become popular in the 1980s when fiscal constraints of public budgets became more pronounced. Additionally, ideological shifts contributed to a tendency away from universalistic towards targeted social policies, claiming that public resources should be devoted only to the most vulnerable members of society (Mkandawire, 2005). Consequently, the success of such targeted development programs critically depends on identifying the right beneficiaries. For them, the decision who becomes eligible and who does not is crucial, as it can hugely influence their ability to make a living, receive basic medical care or provide their children with enough food (Daly and Fane, 2002; Gertler, 2004; Schultz, 2004).

In developed countries, this targeting process is usually based on the income and assets of a household. The government analyses current and previous earnings and assesses the value of assets to determine who is eligible for social security. However, as households in developing countries are often self-employed, work in informal sectors or in agriculture, it is often unclear to the government which households are the poorest (Deaton, 1997). Hence, different processes have been put in place to decide which households will become beneficiaries of social programs. Governments can, for example, define simple rules based on the households' demographics or location, or ask local leaders to agree on who should receive the benefits in a community (Coady et al., 2004*a*). Alternatively, governments can also assess the welfare levels of individual households through extensive surveys called means tests.

Another class of methods that governments apply are proxy means tests (PMTs) (Grosh and Baker, 1995). PMTs utilize detailed surveys on consumption for a representative subpopulation to calibrate statistical models that, in turn, predict consumption based on observable household characteristics. This allows governments to avoid conducting the more expensive surveys on consumption for the whole population. The characteristics usually include easily observable, objective information on the household such as the type of dwelling and regional characteristics like the access of the local community to medical services. After developing a precise model, the government only needs to conduct short surveys on observables to obtain consumption estimates for the remaining majority of households. Hence, PMTs represent a compromise, as they empirically target poor households better than simple geographic or demographic targeting and are at the same time cheaper than means tests (Coady et al., 2004*b*).

Yet, PMTs have significant shortcomings which are discussed both in practice and the academic community (Bennett, 2017). One objection is that proxy means tests are difficult to communicate and are thus not always perceived as fair by the respective communities compared to other targeting mechanisms (Alatas et al., 2012). Additionally, PMTs do not target perfectly and current practice may result in 18 to 22 percent of households being misclassified (McBride and Nichols, 2016). Misclassification occurs when a poor household is classified as non-poor by the targeting mechanism, or vice versa. While a variety of statistical approaches can be used for a PMT, program directors mostly rely on ordinary least squares (OLS) to establish correlations between household characteristics and consumption. However, due to their functional form, OLS regressions only establish linear relationships. Thus, they might not be the optimal approach to predict consumption if more complex relationships, such as non-linear ones or interactions of characteristics, are important predictors of consumption. Other statistical prediction methods such as penalized regressions or machine learning techniques could be used to overcome these limitations. Due to the continuous increase in data availability and computation power, these methods now can be applied in econometrics (Varian, 2014), making them promising tools to reduce the misclassification rates of proxy means tests.

The purpose of this thesis is to assess the prediction accuracy of different statistical methods in the context of poverty targeting. We compare four methods in total. While OLS serves as our benchmark, penalized regressions are used as an alternative econometric method. To evaluate the potential of machine learning tools, we also apply neural networks and random forests. Our analysis is conducted using two multi-topic surveys from India and Indonesia, the India Human Development Survey and the Indonesian Family Life Survey. We choose these surveys because they are panels and contain data representative for a large share of the respective population. In addition, India and Indonesia are among the four most populous countries in the world and have undergone rapid growth during the last decades (The World Bank, 2019).

When evaluating the prediction methods, we analyze three different facets. First, there are aggregate misclassification rates. We look at the total share of households misclassified, which we denote as the total error rate. We also differentiate between the share of households that are wrongly included in an anti-poverty program, the inclusion error rate, and those that are wrongly excluded from the program, the exclusion error rate. As including non-poor households represents a misallocation of public funds and excluding poor households contradicts the program's objective, analyzing these errors separately is a necessity. Second, we evaluate how the methods classify different consumption percentiles and whether there are heterogeneous targeting outcomes for important subgroups. Third, we study the stability of the consumption predictions over time. In practice, censuses are only conducted with significant time gaps and the underlying relationships captured by the proxy means test can change as societies develop in the meantime.

Our research relates to the literature on social safety nets, poverty targeting and machine learning. Social safety nets have been implemented for poverty alleviation in developing countries since the 1970s (Litvack, 2011; Subbarao and Smith, 2003). In the following decades, the focus shifted from broader public work schemes or food subsidies towards conditional cash transfers and other targeted approaches (Subbarao and Smith, 2003).

Coady et al. (2004*b*) assess different targeting methods for those programs; another, narrower analysis of proxy means tests is conducted by Alatas et al. (2012). The authors compare the targeting performance and local reception of proxy means tests, community targeting methods and a hybrid version through a field experiment in Indonesia. They find that proxy means tests most accurately identify the poor when using per capita consumption as the measure of poverty. Yet, they state that the differences are small in an economic sense. Therefore, we consider it worthwhile examining whether the misclassification rates of PMTs can be further reduced. McBride and Nichols (2016) explore this idea by applying random forests, a machine learning algorithm, on different USAID data sets and argue that such tools have the potential to improve proxy means tests. While their analysis makes a strong case for further exploring machine learning methods in the context of PMTs, the study falls short on some dimensions. Namely, they rely on comparatively small datasets, consider only one machine learning method and use only a small set of variables for prediction.

With our thesis, we make three contributions to the existing literature. First, we investigate alternative statistical methods for PMTs in a very different setting than McBride and Nichols (2016). With India and Indonesia, we study two large and heterogeneous countries, allowing us to investigate whether the methods discriminate against important subgroups. Also, we use larger data sets and a bigger set of variables. Second, we are, to the best of our knowledge, the first to apply neural networks for proxy means tests. Third, we test the stability of the predictions over time and thus assess the suitability of the different methods for environments in which new consumption surveys cannot be conducted frequently.

Our study reveals that the prediction accuracy of the four methods does not differ strongly. While the neural network achieves the lowest total error rate in most of our analyses, the differences between the methods are statistically significant only in few of them. The differences to OLS are always less than one percentage point and thus economically small compared to total error rates of around 17 percent in the baseline analysis. This pattern holds for the inclusion and exclusion error rates. Looking at the gender of the head of household, urban vs. rural households and households living in different states, we do not find any systematic differences in targeting accuracy either. We confirm these results using a smaller set of variables and a poverty line at half the levels of the official ones.

One important finding is that all methods predict well over time in the fast-growing economies India and Indonesia. Calibrating the methods on the first and predicting on the second survey round increases the total error rates by less than two percentage points, even though the data has been collected more than five years later. However, we cannot distinguish any large differences between the methods in this analysis either.

The rest of this thesis is organized as follows. Section 2 provides background information on poverty targeting, proxy means tests and machine learning. In Section 3, we give a detailed overview of the data sets and the variables selected for the PMT[1]. Section 4 describes our methodology and the application of each method in greater detail, with a focus on the penalized regressions, random forests and neural networks. Section 5 presents our main results and is followed by robustness checks in Section 6. Section 7 contains motivation, approach and results for our investigation of stability over time. A discussion follows in Section 8, before Section 9 concludes.

2 Background

2.1 Measuring Poverty

Although difficult, measuring living standards and welfare is integral for social policies and informs the debate on poverty and inequality. Consequently, the discussion among economists how to measure poverty has been going on for decades. While there seems to be agreement that poverty is not unidimensional, evaluating social policies often requires a single metric (Atkinson and Bourguignon, 1982). Thus, the two monetary indicators income and consumption have been suggested as they aggregate different dimensions such as the ability of a household to buy enough food or fund the education of children (Gillis et al., 2001). Which of the two to use has been subject to another debate as it depends on the development status of the respective country.

Deaton (1997) suggests using consumption as the welfare measure in developing countries as theory implies that an agent's consumption is her smoothed representation of lifetime income. Income might accrue at certain periods within a year, for instance in the season when farmers sell their crops. Thus, using income from that period would overestimate their welfare, while taking income from other periods would underestimate it. Consumption, on the other hand, would likely not differ too much between the periods and is therefore a less volatile measure of actual living standards. Additionally, rural households in developing countries tend to source large parts of their incomes from self-employment which makes them hard to observe from a practical standpoint.

The question follows what constitutes the consumption of a household. Deaton and Zaidi (2002) refer to this as the consumption aggregate and split the components into four classes. The first of these classes are food items which are usually split in purchased and non-purchased food, the latter representing homegrown food. The second class refers to non-food consumption such as expenditures dedicated to health and education, but also daily purchases like clothing, petrol and recreational expenses. Consumer durables such as appliances or vehicles represent the third class in the consumption aggregate, housing costs the last.

The data to construct these consumption aggregates is usually collected through house-

[1] In the regression context, these are called independent variables or regressors. This differs from the machine learning literature where they are called predictors or features (Varian, 2014). For consistency, we use the term (independent) variables throughout this thesis.

hold surveys. Grosh and Glewwe (2000) have compiled a guideline for designing household surveys in developing countries, building on extensive experiences with the World Bank's Living Standard Measurement Surveys and other multi-topic household surveys in the developing world. But even with a perfectly designed survey, misreporting will occur. One reason for this is limited ability to recall (Deaton, 1997), while conscious misreporting poses a problem as well. Martinelli and Parker (2009) investigate the extent of misreporting in household surveys used in Mexico's *Oportunidades* program and find systematic and widespread misreporting, which gets worse with increasing program benefits. Additionally, they uncover systematic overreporting of goods and home characteristics linked to social status. Nevertheless, gathering consumption data through household surveys and interviews is best practice to assess living standards in the developing world (Deaton, 1997).

With consumption as the welfare metric, measuring poverty still is not straightforward. Notably, one can adopt notions that define poverty either in relative or absolute terms. Sen (1973) has argued to perceive poverty as the inability to properly function in society which implies a relative assessment of consumption. Alternatively, absolute definitions of poverty have been adopted by many countries such as India or the United States (Orshansky, 1963; Subramanian and Deaton, 1996). In developing countries, paying for the minimum nutritional intake will constitute an important part of that consumption threshold. However, these thresholds should be viewed with caution, as they are unlikely to be measured precisely and as it is debatable whether a sharp cut-off between the poor and non-poor exists (Atkinson, 1987; Deaton, 1997).

Defining and measuring poverty is not a focus of our thesis, but it is important to keep the ambiguity and limitations of these definitions in mind when looking into anti-poverty programs. Even in a world where we would be able to perfectly record unbiased consumption data through household surveys and were entirely certain about the position of a sensible poverty line, we would not be able to use surveys to accurately assess the poverty status of each single household. Interviewing households represents an exercise that cannot not be undertaken for the whole population in practice as it is expensive and time consuming. Hence, approaches have been developed and implemented that are more feasible in such contexts.

2.2 Anti-Poverty Programs and Targeting Mechanisms

Despite disagreement on the exact definition of poverty, economists and politicians agree that there is still enormous poverty in the world, particularly so in large parts of Africa, South America and Asia. Consensus is growing that further economic growth will not be sufficient to eradicate poverty and additional measures are required to tackle the issue (Hanna and Olken, 2018). Governments can employ different kinds of anti-poverty programs, some of which are universalistic, others targeted. Universalistic programs such as food subsidies or schooling programs fight poverty indirectly through better nutrition or education (Daly and Fane, 2002; Duflo, 2001), while targeted anti-poverty programs tackle poverty by directly supporting the households in need and are often designed as national transfer schemes. However, this comes with the challenge of correctly identifying those households. This can be illustrated with Brazil's *Bolsa Familia* program, a conditional cash transfer program, which accounted for 0.7 percent of the national spending in 2009 (calculations based on Berg, 2010; World Bank Group, 2018). At the same time, Dutrey (2007) shows that the undercoverage ratio, the share of poor households that have not been reached by the program, amounts up to 73 percent. As the targeting mechanism has a big impact on the effectiveness of any anti-poverty program, it needs to be chosen carefully.

Coady et al. (2004*b*) classifies these mechanisms into three broader categories. First, there is targeting that involves the assessment of individual households, which can be done with means tests. They assess through detailed consumption or income surveys whether a household is eligible for an anti-poverty program. To reduce administrative costs, proxy means tests have been developed in which a smaller number of household characteristics is collected in order to assess program eligibility. Also, community targeting falls into the category of individual household assessments. In this case, members of the community rank all households directly by compiling a poverty-ranking based on which program participation is decided. Another class of targeting approaches is categorical targeting such as geographic targeting where households are selected based on poverty maps. Alternatively, sometimes simple demographics like age, gender or ethnic origin are used to target households. This is called demographic targeting. Finally, self-targeting methods constitute an own category. They involve a self-selection into a program, which might require a household member to go to an office and file an application (Alatas et al., 2016). In practice, these different targeting mechanism are often combined.

Program directors face a trade-off between costs and precision when deciding on which targeting mechanism to use. While precision refers to identifying poor households correctly as poor or non-poor and is easily understood, costs require a more detailed consideration. Bennett (2017) breaks them down into four types. First, there are design costs that are incurred to prepare, develop and test a targeting method. Second, there are operational costs once the targeting system is up and running, like staffing or communication. Third, there are external costs such as transportation costs to travel to an office and finally, there are opportunity costs that are incurred when filing an application. Usually, the better a method is at targeting the poor, the less practicable and more costly it is (Bennett, 2017). For example, geographical targeting is conducted very cheaply, as program directors only need to select the target areas. However, as everybody in this area will benefit, likely also non-poor households are included in the program. On the other hand, means tests are expected to be very precise, but at the same time costly to conduct.

2.3 Proxy Means Tests

As they offer a good balance between precision and costs, famous anti-poverty programs such as Mexico's *Oportunidades* use proxy means tests to target their beneficiaries. The general idea of a proxy means test is to indirectly assess the household's means through a number of observable characteristics that are indicative of the household's economic welfare.

The mechanism can be explained in three steps. First, the relationships between household characteristics and per capita consumption levels are captured via statistical methods for a small, representative sub-sample of the population. As of today, this is done mostly by running stepwise OLS regressions, often at the state level to capture different local effects (Coady et al., 2004*a*). In a first regression, all available household characteristics are used as independent variables and regressed on per capita consumption. Then, all variables that are not significantly correlated with consumption are removed from the equation and the regressions are re-run. Second, using these patterns, consumption for the remainder of the population, the out-of-sample households for which only the characteristics are collected, is predicted. In the final step, households get classified as poor or non-poor using official poverty lines. Thus, while proxy means tests rely on linear regressions to predict consumption, a binary decision on eligibility is made eventually.

However, this approach still does not perfectly classify households. To assess targeting performance some authors look at the total error rate, defined as the ratio of all households that are misclassified (Alatas et al., 2012). A household counts as misclassified if its per capita consumption level is below the poverty line and it was not deemed eligible or vice versa. It is important to understand that predicting consumption well on average correctly is not sufficient, as those predictions are used to determine legal entitlements that have crucial economic impact for an individual household. There are other measures for targeting performance that we will not discuss in our thesis. We recommend the World Bank report by Coady et al. (2004*a*) for a detailed discussion on these measures and we elaborate further on our metrics in Section 4.

Besides misclassification, there are other shortcomings of proxy means tests. One important drawback is that econometric methods must be used to predict consumption. This makes it difficult to communicate or justify in detail how eligibility decisions are determined, which might inhibit local participation and acceptance of the targeting method (Cameron and Shah, 2013). Even though less so than means tests, proxy means tests still require significant administrative sophistication and capacity, making them not suitable to every type of environment (Mkandawire, 2005).

2.4 Machine Learning

Nowadays, machine learning tools are widely used for prediction purposes in our daily life. Search queries in the internet, language translation as well as image recognition are largely driven by machine learning techniques (LeCun et al., 2015). The statistical methods behind those tools have already been developed and applied for different research purposes in the 20th century. For example, Tu (1996) discusses the advantages of neural networks over logistic regression to predict medical outcomes and Desai et al. (1996) find that neural networks outperform linear scoring models in classifying bad loans for credit unions. Since then, exponentially more data has become available for analysis and better hardware allows researchers and practitioners to train more complex models on larger data sets (Varian, 2014).

As economic research is often concerned with inference rather than prediction tasks, the adoption of machine learning methods has been relatively slow in the field (Einav and Levin, 2014). For instance, approaches that combine the high prediction power of machine learning with econometric methods to determine causality are being developed. Hartford et al. (2017) show how neural networks can be used in the first stage of instrumental variables regression to minimize the counterfactual prediction error of the first stage. Mullainathan and Spiess (2017) provide a more general overview on how machine learning methods can be used to supplement the econometricians' toolbox, as does Varian (2014). As one example, he mentions a situation where there are more potential predictors than appropriate for estimation. In that case, the variable selection can be efficiently conducted by machine learning tools.

For targeted anti-poverty programs, the major statistical challenge is to determine which household is poor, based on individual household characteristics. As outlined in the previous paragraphs, this usually is a task of predicting household consumption, making it a promising field to explore the use of machine learning tools. Both their ability to capture non-linear relationships and their strong performance in predicting out-of-sample (Mullainathan and Spiess, 2017) makes them particularly well suited for the task at hand. McBride and Nichols (2016) explore this for random forests using data sets from Bolivia, Malawi and East Timor and find that they can increase the precision of poverty targeting compared to more traditional approaches.

3 Data Strategy

3.1 Data Sources

In this thesis, we rely on two multi-topic surveys as our main data sources, the India Human Development Survey (IHDS) and the Indonesian Family Life Survey (IFLS). Using these specific data sets entails significant advantages. Both sets have panel character and we analyze data from two rounds of each survey. They are also publicly available and contain a large number of observations. And finally, India and Indonesia offer characteristics that make them an optimal fit for our research question. With the BLT, Indonesia has already designed a social policy based on proxy means tests and India's opposition Congress Party recently suggested a targeted minimum income for the poor during their campaign for the national elections (Biswas, 2019). Both countries are democracies with big differences in income across states and provinces (Asra, 2000; Deaton and Dreze, 2002). Additionally, they have experienced high average GDP growth rates of 5.5 (Indonesia) and 6.7 (India) percent in the last decade and analyzing how the different methods perform over time therefore becomes crucial to assess the potential and limitations of proxy means tests in practice (The World Bank, 2019).

The IHDS is supervised by the University of Maryland and India's National Council of Applied Economic Research. It compiles information on 41,488 households in the first survey round from 2005 and 41,491 households in the second survey round from 2011/2012 (Desai et al., 2010; 2015). Most of the households were interviewed in both survey rounds and thus represent a panel, while 2,134 households were included as replacement for households where contact was lost. The sample of households is nationally representative and covers all states and union territories of India besides Andaman and Nicobar Islands, and Lakshadweep. The data is collected through two one-hour interviews in each household and covers topics like health, education, employment and socio-economic status. Additional information on the village such as the availability and quality of schools and medical facilities is compiled as well.

The IFLS is organized by the RAND corporation and the Center for Population and Policy studies of the University of Gadjah Mada. We utilize the IFLS4 survey from 2007/2008 providing information on 11,631 households and the IFLS5[2] survey from 2014/2015 containing 14,056 households. 9,744 households are included in both survey rounds (Strauss et al., 2009; 2016). The IFLS rounds are based on a sample from 1993, representative for 83 percent of the Indonesian population, and covering households living in 16 out of the 26 provinces of the country[3]. Completing the household surveys often takes several hours but can mostly be completed in one visit, sometimes more than one visit is required. The Indonesian data set also covers information on housing condition, household economy and local characteristics.

3.2 Data Preparation

Given both data sets contain several hundred variables, we must choose those that we consider most useful to predict consumption before running the proxy means tests. Our selection of variables follows the literature and includes most of the variables that are covered by Alatas et al. (2012), Brown et al. (2016) or McBride and Nichols (2016). Also, we rely on Glewwe et al. (1989) who have explored which variables are best suited to predict consumption.

The variables we choose can be summarized in four dimensions. First, demographics provide basic information on household consumption. For instance, the number of young children in a household gives information on the amount of people that need to be supported by the income earners. Second, housing conditions are good predictors for consumption levels as the quality of the floor or whether a private toilet exists proxy the household's welfare. Third, assets and household financials are highly relevant to predict

[2] For our purposes, we will refer to IFLS4 as the first survey round and to IFLS5 as the second survey round to have a notation that is easily interpretable and consistent with the Indian data sets.

[3] This representative sample was drawn for IFLS1. Since then, resampling has been conducted, split-off households (e.g. children) have been included in the sample as well and some provinces have been split. As a result, the IFLS5 data includes households from 23 out of Indonesia's now 34 provinces.

Table 1: Overview of Selected Variables

Country	Data Set	Year	Obs.	Category	Selected Variables	Count
India	IHDS1 IHDS2	2005 - *First Round* 2011/2012 - *Second Round*	41,488 41,491	*Demographics*	# of persons, persons_sq, dep. ratio, age hh head, age_sq, education hh head, education oldest adult, married, widow, caste, # of children, children in school, disabled hh members, gov relation, sex hh head	42
				Housing	ownership type, # of rooms, electricity, solid floor, solid roof, private toilet, type water access, type kitchen, solid wall, wood for cooking	
				Assets	size of loans, farm income, # buffalos type occup. hh head, # cows, fridge, motorbike, fan, telephone, cellphone, # of persons farming, bike, tv	
				Local features	urban, region, state, access to doctor	
Indonesia	IFLS4 IFLS5	2007/2008 - *First Round* 2013/2014 *Second Round*	11,631 14,056	*Demographics*	# of persons, persons_sq, sex hh head, age_sq, # of children, children in school, dep. ratio missing, level education hh head, level education oldest male, age hh head, dep. ratio, marital status	35
				Housing	type ownership, solid floor, solid wall, type toilet, type water used, type cooking fuel, size per capita, electricity	
				Assets	tv, fridge owned, fridge used, loans, type occup. hh head, work type hh head, farm activity, size of farm	
				Local features	urban, region, province, clinic distance, aware of clinic, posyandu distance, aware of posyandu	

consumption. For example, we include a dummy on whether the household owns some land for farming (Narayan and Yoshida, 2005). Also, the ability to purchase assets such as a motorbike or a fridge points towards a higher consumption level of a household. Finally, local characteristics provide insights about the healthcare infrastructure which is expected to be better for high-income areas.

After selecting the variables, we merge individual survey data with household survey data. In order to construct a consumption per capita variable for the Indonesian data set, we transform all relevant expenditures into monthly per capita values and aggregate them. For India, per capita consumption values are already in the data set. This includes food as well as non-food expenditures and education expenses. To make the remaining variables usable for analysis in a regression or the machine learning tools, we process categorical variables. As one example, for India, we find a categorical variable containing information on the ownership type of the accommodation, containing three different values. Either the dwelling is owned, rented or the ownership situation is unspecified which we translate into three dummy variables. Finally, metrics like the dependency ratio are not available in the raw data and thus we compute them. This process results in 42 characteristics for India and 35 characteristics for Indonesia which are depicted in Table 1.

There are only few data points missing in both data sets which we deal with following Papageorgiou et al. (2018). In a first step, we check each variable for a sizeable number of missing values. For variables where this is the case, we assess the reason why the data is missing. In the Indian data set, for many assets the value "valid blank" is recorded. For example, for the variable capturing the number of cows, we interpret this as households not engaging in farm activity and thus owning no cows. Another example is represented by the dependency ratio in the Indonesian data sets. If a household has no member that worked in the last 12 months, the denominator equals zero and the dependency ratio cannot be calculated. To capture this without dropping the observation, we assign the household the mean and create a new dummy variable that captures whether the dependency ratio was set manually to be the mean or not. In this way, we do not skew our regressions and do not lose observations. Finally, we drop all observations for which we do not have information on consumption.

To categorize households into poor and non-poor, we use the official, national poverty lines. In the case of India, we rely on the publications of the Planning Commission of the Indian government which defines the poverty lines (Government of India Planning Commission, 2009; 2013). Poverty lines are reported separately for urban and rural households. For Indonesia, we refer to Priebe (2014) who published an extensive review on the poverty measurement by Indonesia's statistical agency BPS.

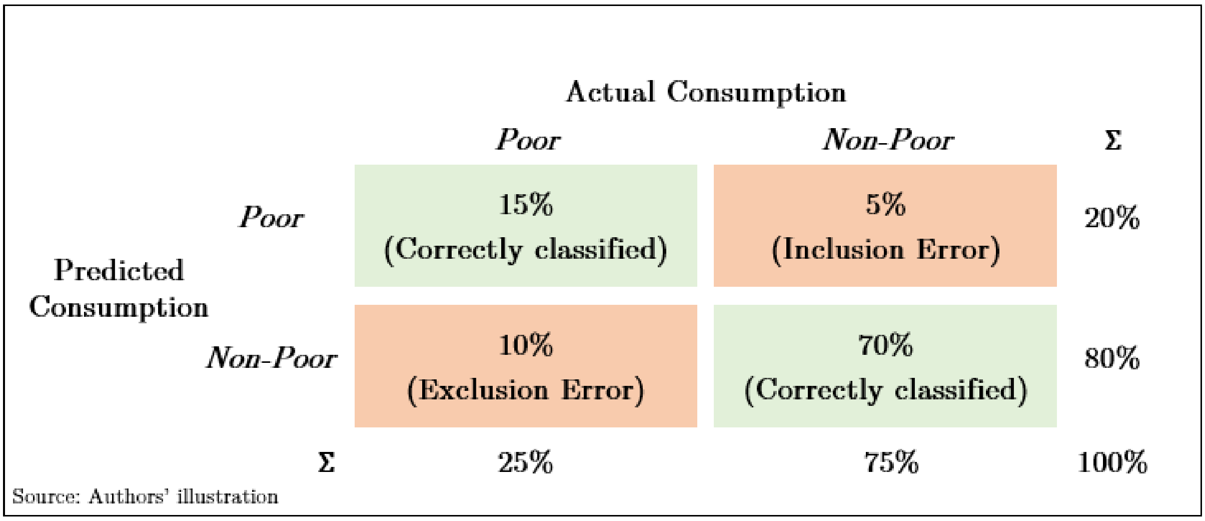
4 Methodology

The aim of this thesis is to evaluate different methods in econometrics and machine learning in their performance in poverty targeting. The two econometric methods applied are OLS, which serves as the benchmark used in practice, and penalized regressions. The machine learning tools that we explore are neural networks and random forests. We motivate the selection of each method in their respective subchapters.

To simulate the setting in practice, we randomly split our data sets into a training and a test set, which both contain household characteristics and consumption data. Only the training set is then used for the calibration of the models, while we use the test set to assess the out-of-sample targeting performance of the different methods. Throughout this thesis, we follow Friedman et al. (2001) and chose a 75/25 split, meaning we define a training set with 75 percent of all households and a test set with 25 percent of the households. To ensure that the results are unskewed, we split the data before the calibration process and assess the targeting performance of the methods with the same test set. Table A1 and A2 in the Appendix show the means for the first round data sets as well as a difference-in-means test for the training and test set and confirm that they are balanced. The data is also balanced for the second round which we do not report here.

We follow the procedure for the proxy means test as described in Section 2.3 to assess the targeting performance of the models. We mainly analyze the share of households in the population that are misclassified, which we illustrate with Figure 1. Based on the actual consumption of the households and the predictions by the different models, we categorize the households in three different categories. Households with a predicted consumption above the poverty line whose actual consumption is also above the poverty line are correctly classified. In the example, this refers to 70 percent of the households in the green box in the lower right. Also correctly classified are households that actually poor and that are also predicted with a consumption lower than the poverty line, represented by the 15 percent in the green box in the upper left.

Figure 1: Illustration of Error Rates



		Actual Consumption		
		Poor	*Non-Poor*	Σ
Predicted Consumption	*Poor*	15% (Correctly classified)	5% (Inclusion Error)	20%
	Non-Poor	10% (Exclusion Error)	70% (Correctly classified)	80%
	Σ	25%	75%	100%

Source: Authors' illustration

Households that have an actual consumption below the poverty line but are predicted to have a consumption above, represent an exclusion error. Based on their actual status, they should have become eligible for the program, but the prediction tells us otherwise. In the example, this refers to the 10 percent of the households in the red box in the lower left. Finally, households that are actually non-poor but are predicted to have a consumption level below the poverty line, represent an inclusion error. These households should not have become eligible for the program, but they were predicted to be eligible. In the example, this refers to the 5 percent of the households in the red box in the upper right. Finally, we define the total error rate as the sum of the two rates, which is the total share of households that are misclassified by a method. In the example, this number would amount to 15 percent. For each method, we choose the model specification that yields the lowest total error rate on the training set and then calculate our evaluation metrics. In a more extensive analysis, we compare the error rates across major subgroups such as different regions or urban vs. rural households. We provide an overview of all analyses conducted in this thesis in Table 2.

In Sections 4.1 to 4.4, we motivate each method, describe how it establishes the relationship between household characteristics and consumption and how it is optimized to

Table 2: Overview of Analyses

	Dimensions			
Analysis	Data Set Round	Variable Set	Poverty Threshold	Section
Baseline	First	Long (∼40 characteristics)	Nat. Poverty Lines (NPL)	5.1
Second Round	Second	Long (∼40 characteristics)	Nat. Poverty Lines (NPL)	5.2
RC: Half Poverty Line	First	Long (∼40 characteristics)	50% of NPL	6.1
RC: Short Vector	First	Short (∼15 characteristics)	Nat. Poverty Lines (NPL)	6.2
Time Stability	First & Second	Long (∼40 characteristics)	Nat. Poverty Lines (NPL)	7

Note: RC: Robustness Check

improve the targeting performance. Then, in Section 4.5, we discuss differences between the methods.

4.1 Ordinary Least Squares (OLS)

Proxy means tests based on OLS regressions have been used widely in practice and consequently represent our benchmark we compare all the following methods against (Brown et al., 2016). For OLS, we follow the estimation approach of Alatas et al. (2012). That is, we bundle the training set of a country in regional units and run the OLS regression in equation 1 for each regional unit using all the input variables from Table 1 in the first step. Then, all variables whose coefficients are not significant on a 10%-level get removed and the regressions are re-run with the smaller set of input variables in a second step. Finally, we predict the consumption of the households in the test set using the regional-specific second-step coefficients. The results for the second-step regressions for the baseline analysis are reported in Tables A3 and A4 in the Appendix.

$$log(\text{cons}_i) = \alpha + \beta * \mathbf{demographics}_i + \gamma * \mathbf{housing}_i + \delta * \mathbf{assets}_i + \theta * \mathbf{local\ features}_i + \epsilon_i \tag{1}$$

4.2 Penalized Regressions (PR)

Penalized regressions differ from OLS by adding penalizing terms to the minimizing function of the model. The penalizing terms bias the coefficients of the different variables towards zero, thereby regularizing the model to reduce the influence of possibly unnecessary variables. As Mullainathan and Spiess (2017) state, regularization can help to decrease overfitting and improve out-of-sample predictions of statistical models. Hence, we explore the possibility that penalized regressions outperform OLS methods on our data sets.

Varian (2014) illustrates the technical details behind penalized regressions intuitively. Equation 2 denotes the penalizing term that is added to the formula which OLS regressions use to minimize the sum of squared residuals.

$$\lambda \sum_{p=1}^{P} [(1-\alpha)|b_p| + \alpha|b_p^2|] \tag{2}$$

α represents the so-called hyperparameter that can be varied between 0 and 1 to improve the prediction performance of the method. $\alpha = 0$ leads to the "least absolute shrinkage and selection operator" (LASSO), while the case $\alpha = 1$ represents the so-called ridge regression. If α is between 0 and 1, the regression is called an elastic net. All three variations of the penalized regression class lead to smaller coefficients for many input variables, with LASSO being the strongest version, setting many coefficients to zero. λ, on the other hand, defines the weight of the penalizing term, with $\lambda = 0$ representing a standard OLS regression. Hence, there are two hyperparameters within the penalized regression method that can be optimized to obtain the best targeting performance.

Hyperparameters are high-level parameters that are defined before applying a method on a data set and remain constant throughout the development of the corresponding model. Each of the methods we use has a set of hyperparameters, which in turn lead to different predictions. Whilst the respective process of optimizing the hyperparameters is described in the following paragraphs, it is essential to note that this optimization is only conducted analyzing the prediction performance of the models on the training set, as we would otherwise introduce bias. For the optimization, we follow a standard approach in machine learning that is suggested by Friedman et al. (2001). We split the training set randomly into a new, smaller training set, two thirds of the size of the original one. This way, half of all households are in this new training set and the remaining third of the initial training set becomes the validation set. We conduct the hyperparameter tuning for the random forests and the neural networks analogously. This procedure promises models that are well specified for out-of-sample predictions (Varian, 2014).

To optimize the penalized regression method, we apply Stata's *cvlasso* package by Ahrens et al. (2018) and grid search on the small training set. The package automatically optimizes the λ parameter using cross-validation[4]. To optimize the remaining hyperparameter α, we conduct grid search, trying out all α parameters between 0 and 1, using 0.1 increments. In a second step, we evaluate their performance on the validation set and pick the α that yields the lowest total error rate. To obtain the final predictions for the penalized regression method, we then train the model with this optimal α on the full training set and apply it on the test set. The respective optimal levels of α for each analysis are recorded in Table A7 in the Appendix.

4.3 Random Forests (RF)

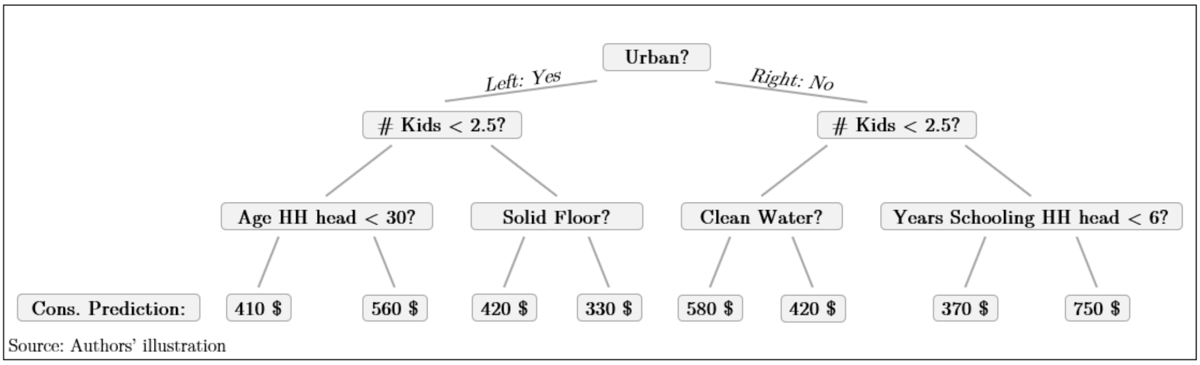
Random forests are an extension of regression trees and a popular machine learning algorithm. Varian (2014) provides a concise summary on the advantages of random forests for classification purposes and attributes a particularly good out-of-sample performance for non-linear data to random forests, which is why we consider it a method worth exploring for our research question. Additionally, McBride and Nichols (2016) have already demonstrated that they perform well for prediction tasks in the context of proxy means

[4] An explanation of cross validation can be found in Section A.1.2 in the Appendix.

tests.

A random forest is a recursive splitting algorithm and consists of many regression trees. The typical procedure to grow a forest can be described in a few steps. First, a random sample of observations is bootstrapped from the training set and used to grow the first regression tree. At each node of the tree, a random sample of variables is selected from the full set and the algorithm then determines for each variable the split point at which the summed squared distance between the mean predicted outcome and the mean actual outcome is minimized. Through this, non-linear relationships can be captured. The variable with the split point that leads to the best predictions gets implemented into the tree. This process is repeated at each node until the tree is fully grown. We show an example of a regression tree in Figure 2. These steps are repeated until the forest is fully grown. As each tree is based on different subsets of the population and looks different from each other, random forests are expected to capture heterogeneous outcomes quite well. To calculate the prediction of the forest, the average of the predictions of all trees is taken. For a more detailed mathematical background, we recommend the original paper by Breiman (2001) or the textbook by Friedman et al. (2001), which covers both regression trees and random forests in particular.

Figure 2: Example of a Regression Tree



Source: Authors' illustration

We implement the random forests in Python using the *RandomForestRegressor* from the *skicit* library (Pedregosa et al., 2011). There is a small number of hyperparameters that are most relevant to optimize the model, one of them being the maximum size of each tree. While larger trees are more precise, they tend to overfit if they memorize individual outcomes rather than patterns. Yet, if the trees are too short, they may not capture enough information to detect a pattern. For a full list of hyperparameters we optimize over, see Table A8 in the Appendix.

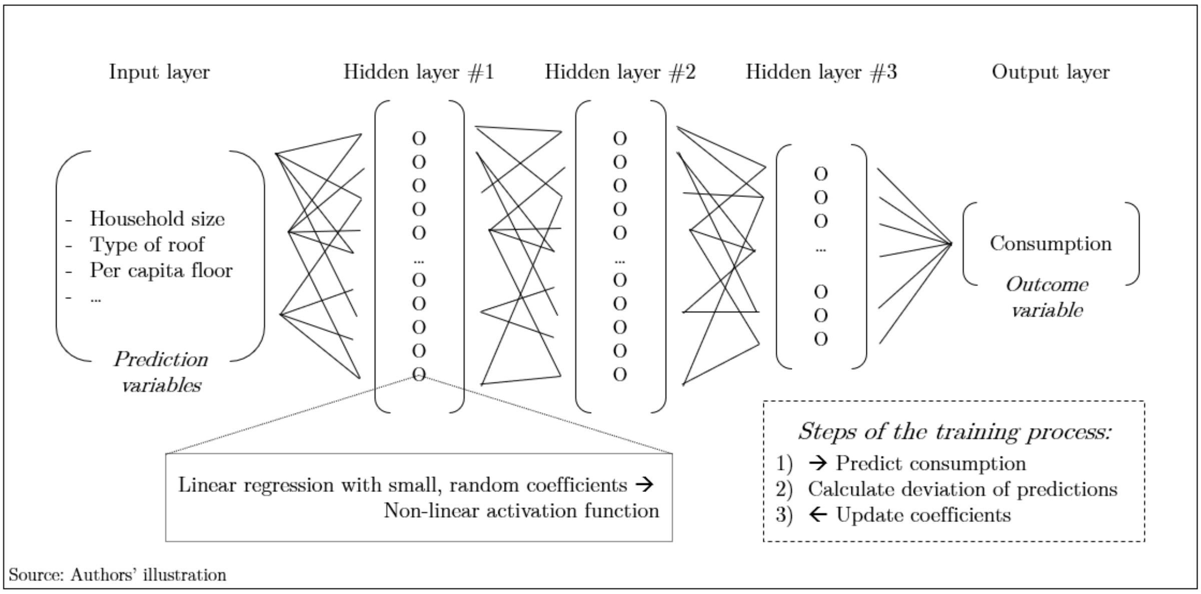
We follow the same approach of using a smaller training and a validation set, as described in Section 4.2 on penalized regressions. However, as there is not only one, but many hyperparameters to tune, finding the optimal set of hyperparameters becomes more complex. This optimization process can be conducted in different ways. While trial-and-error, where the hyperparameters are picked based on the researcher's instincts, can work well, there are more systematic approaches (Bergstra and Bengio, 2012). Especially grid search, which we use for penalized regressions, and random search have become popular. In the latter case, the researcher defines a search space for each hyperparameter and the algorithm randomly picks a combination of hyperparameters and evaluates their performance. However, as Francois Chollet (2018) suggests, we apply the *hyperas* package in Python, an algorithm that optimizes hyperparameters based on trees of Parzen estimators for both the random forests and the neural networks. Table A8 in the Appendix also shows the exact specifications of the random forest models that yield the lowest total error rates on the validation set for each data set. The respective models were then applied on the test sets to predict consumption and calculate the targeting accuracy.

4.4 Neural Networks (NN)

Neural networks are another class of machine learning algorithms widely used in many academic and professional applications due to their flexibility and predictionary power (Haykin, 1994). In this thesis, we use fully-connected, feed-forward neural networks to predict consumption. Desai, Crook and Overstreet Jr (1996) show, using household characteristics such as age and type of ownership, that those kinds of networks can perform better than linear models for building credit scoring models.

Figure 3 represents the rough conceptual architecture of such a fully-connected neural network. Like a linear regression, neural networks start off with the input variables in the form of a vector and connect them to the outcome variable. But instead of determining the coefficients of the input variables directly by minimizing the squares of the prediction error with a linear function, the input variables in a neural network pass through several so-called hidden layers. In each of these layers, there are a number of so-called neurons which represent the outcome variable of their own regression. These neurons are connected to all input variables of the previous layer through a combination of a linear regression and a non-linear activation function. For example, a neuron in the first hidden layer is connected to all household characteristics and simultaneous represents an input variable for all neurons in the second hidden layer. The neurons in the final hidden layer are thus the input variables to predict our outcome, per-capita consumption. Due to this structure, all input variables, neurons and the outcome variable are connected by regressions and their respective coefficients.

Figure 3: Conceptual Structure of a Neural Network



Source: Authors' illustration

Once the architecture of the neural network is defined, a training epoch can begin. Initially, all weights get randomly set to be small but non-zero. Then, the household characteristics of a set of observations from the training set are used to create initial predictions for the households' consumption levels. A loss function such as mean squared error is applied afterwards to calculate how far off those initial predictions are. In the next step, called backpropagation, the derivatives of the loss function with respect to the weights are calculated. In the last step of such a training epoch, the initially randomly assigned weights get updated using those derivatives to improve the predictions in the next training epoch. Intuitively, the network checks in which direction the weights need to change to decrease the loss function. The whole training process is comprised of several training epochs and the input variables get normalized to ensure an effective training process. We also apply cross validation for the neural networks. One important attribute of them is the combination of linear regressions and non-linear activation functions which allow them to capture relationships between variables that cannot be represented by a linear function. For more details on the mathematical architecture behind neural networks, we again refer to Friedman et al. (2001). Additionally, LeCun et al. (2015) provide a concise overview on neural networks and why they have become a popular machine learning method for prediction and classification purposes.

Similar to other estimation approaches, neural networks face the challenge of overfitting. For example, like the size of a tree in a random forest, the number of hidden layers and neurons influences the model's tendency to overfit. Large neural networks with many hidden layers and neurons can fit the training data almost perfectly, but they tend to memorize the results rather than detecting the patterns. To find the set of hyperparameters that gives us the best out-of-sample prediction performance, we follow the same approach used to optimize the random forests. This means utilizing the *hyperas* package and a validation set. The selected set of hyperparameters and the corresponding model get then applied on the test set to evaluate its out-of-sample performance just as for the other methods. For each analysis, we report the optimal set of hyperparameters in Table A9 in the Appendix.

4.5 Comparison of Chosen Methods

Above, we have motivated the choice of each single method. Yet, it is important to understand how the methods compare to each other and what advantages and disadvantages they might imply. A first aspect already mentioned is the ability to capture non-linear relationships and interactions between variables. Ordinary least squares and penalized regressions are bound to the functional forms specified by the researcher. Given it is not clear ex-ante which household characteristics have a potential non-linear relationship with consumption or should be interacted, a linear specification without interactions is chosen and thus these methods are not able to capture any other effects. Should there be any non-linear relationships between consumption and household characteristics, thus, we would expect random forests and neural networks to perform better, as they can detect such effects without the need of pre-specification by the researcher (Chollet, 2018; Varian, 2014).

Second, the methods strongly differ in terms of complexity and required computation power. An advantage of OLS is its easy implementation in statistical tools like STATA and little computational requirements due to simple matrix calculations. This, in turn, limits the researcher's degrees of freedom which are restricted to the confidence level of the variables selected for the second step in our case. Penalized regressions require more sophistication and need to be optimized by the researcher. They can still be implemented through STATA but require more computation power than OLS when choosing the optimal parameter values and performing automatic variable selection (Ahrens et al., 2018; Varian, 2014). Random forests and neural networks require programming skills and are often implemented through R or, as in our case, Python (Chollet et al., 2015). Also, the mathematical foundations of the algorithms are more complex (Breiman, 2001; Friedman et al., 2001; LeCun et al., 2015). An extensive optimization process is needed to find the best model, increasing the degrees of freedom of the researcher (Chollet, 2018).

A disadvantage of OLS is the lack of regularization (Zou and Hastie, 2005). Given regularization decreases the method's tendency to overfit in a prediction task, penalized regressions, random forests and neural networks are generally well-suited to perform out-of-sample prediction tasks, allowing the researcher to reduce overfitting by adjusting the model parameters.

Finally, there are several aspects distinguishing random forests from neural networks. The former are based on the easy-to-visualize concept of a regression tree and thus their predictions can be well communicated to readers familiar with decision trees which are applied in many different areas in economics. Also, they are less computationally extensive than neural networks (Friedman et al., 2001). On the other hand, as Chollet (2018) states, neural networks have recently replaced support vector machines and tree-based algorithms for many prediction tasks and distinguish themselves from other machine learning tools by offering a layered representation of the data, which is where the term deep learning stems from (Chollet, 2018). He argues that by using several transformations of the data, neural networks can solve complex problems more effectively than other statistical tools.

5 Main Results

5.1 Baseline Results (First Round of Each Survey)

Table 3 depicts the error rates of the selected models for the baseline analysis. Segment (1) captures the total error rate, the share of households that have been misclassified in the respective test set, as well as the corresponding differences to OLS and p-values. Segment (2) and (3) depict the exclusion and inclusion error rate respectively. The pattern is the same across all models. For both data sets, the total rate of households being misclassified lies between 16 and 18 percent, which is comparable to the results of McBride and Nichols (2016). Around two thirds of this total error rate represent an exclusion error, meaning that the models categorize households as non-poor, although their actual consumption lies below the threshold.

To analyze whether the rates are different across models, we run the following specification

$$\text{error}_{im} = \alpha + \sum_{m=1}^{3} \beta_m * \text{method}_{im} + \epsilon_{im} \tag{3}$$

where $error_{im}$ is a dummy that equals 1 if household i is misclassified by method m, and 0 otherwise. For the regressions on the inclusion and the exclusion error rate, the dependent variable is adjusted accordingly to equal 1 if the household represents an inclusion or an exclusion error. The βs represent the coefficients of interests, capturing the difference of each method to OLS, which is the reference method in all regressions. Standard errors are clustered at the district level.

For India, all total error rates are significantly different from OLS at least at a the 5% -level. The neural network produces the smallest total error rate, followed by OLS, the penalized regression and the random forest. However, the differences are very small economically. The neural network, misclassifies only 0.63 percentage points less households, which reduces the OLS error rate by roughly 3.6 percent. The other models target slightly worse, the differences however remain economically small.

The results for the Indonesian data set are similar. All total error rates are within 1.2 percentage points, with the neural network again being the most precise. The differences

Table 3: Targeting Error Rates on the Test Set - Baseline

	India								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	17.27	-	-	11.42	-	-	5.85	-	-
Penalized Regression	17.74	0.47	0.003	11.96	0.54	0.000	5.79	-0.07	0.507
Neural Network	16.64	-0.63	0.006	10.22	-1.20	0.000	6.42	0.57	0.003
Random Forest	17.84	0.57	0.040	13.40	1.99	0.000	4.44	-1.42	0.000

Note: The test set comprises 10,371 households. Standard errors are clustered at the district level.

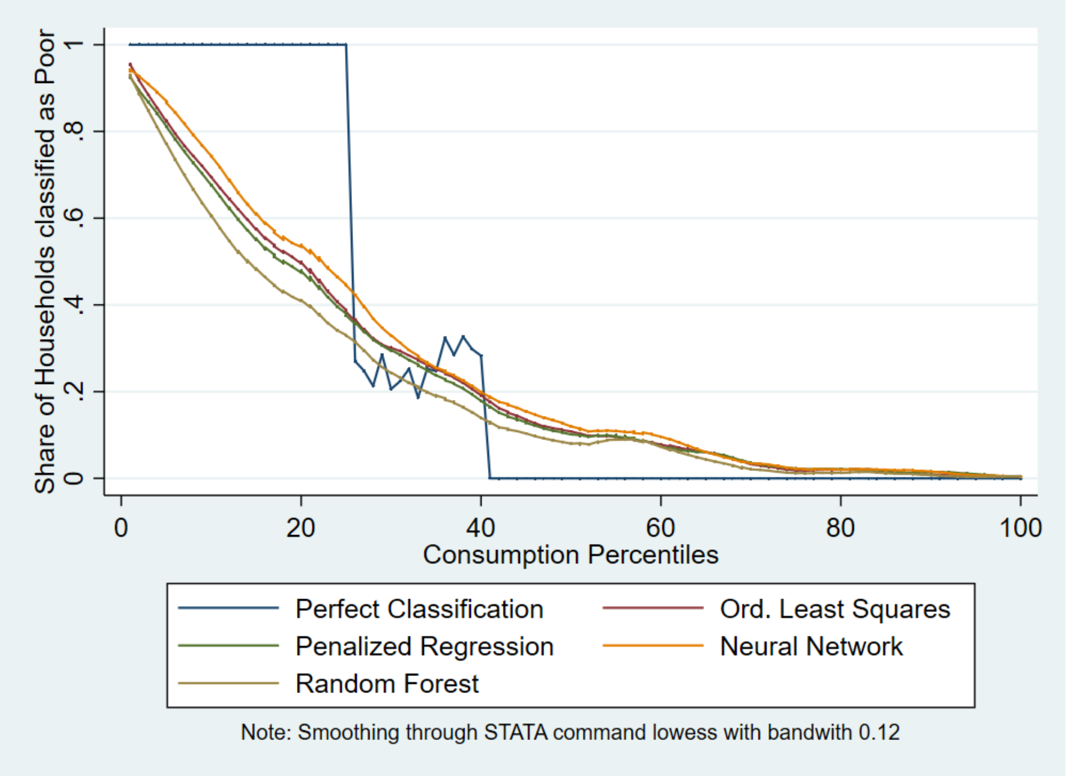
	Indonesia								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	17.02	-	-	13.58	-	-	3.44	-	-
Penalized Regression	16.54	-0.48	0.126	12.69	-0.89	0.001	3.85	0.41	0.052
Neural Network	16.44	-0.58	0.211	11.93	-1.65	0.000	4.50	1.07	0.001
Random Forest	17.64	0.62	0.195	14.48	0.89	0.035	3.16	-0.28	0.397

Note: The test set comprises 2,908 households. Standard errors are clustered at the district level.

are not statistically significant in the Indonesian data set, probably as the data set contains only little more than a fourth of households compared to India. This pattern holds when we differentiate between inclusion and exclusion errors.

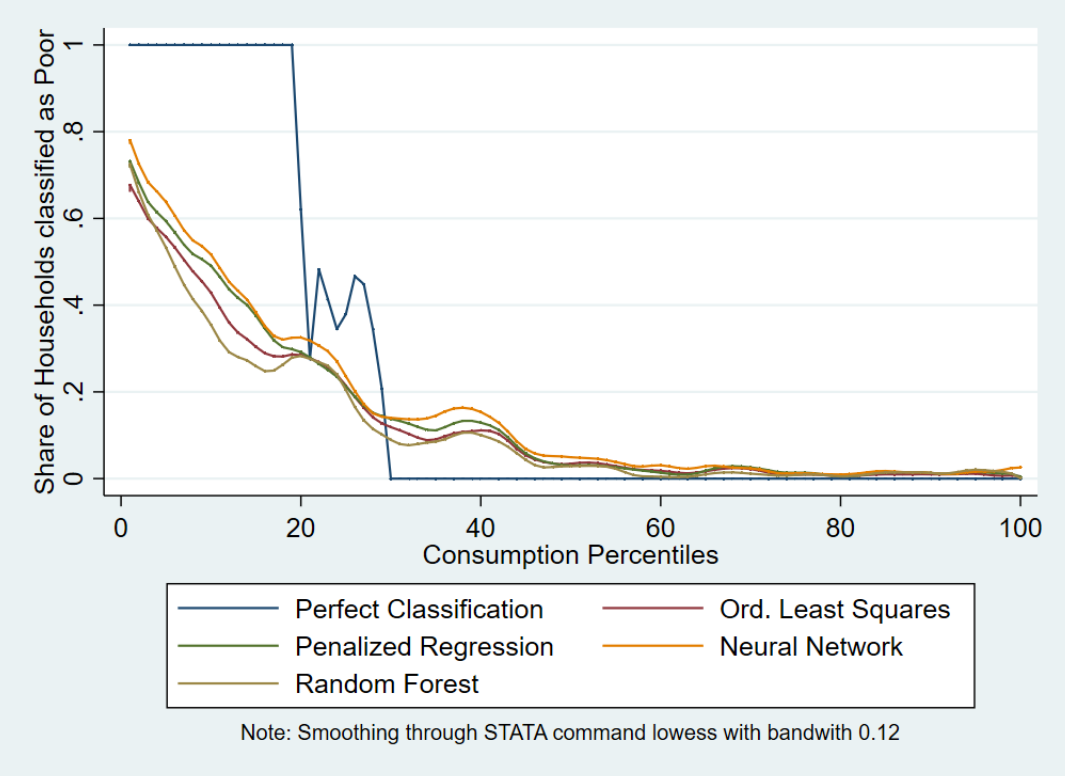
Table 3 shows that there are no economically significant differences in the targeting precision of the different models on aggregate. However, it is important to analyze which household are misclassified. For instance, excluding an extremely poor household will have more severe consequences than excluding a barely poor one. Hence, we analyze how well the different models predict poverty across consumption percentiles. Figures 4 and 5 capture this relationship. On the horizontal axis, we plot all households in the test sets, sorted by consumption and binned into 100 percentiles. The vertical axis depicts the share of households in the respective percentile categorized as poor by the different models. The blue line depicts the perfect classification with all households below the rural poverty line categorized as poor and all households above the urban poverty line categorized as non-poor. The spikes of this blue line stem from differing shares of urban and rural households in the respective bins.

Figure 4: Targeting along Consumption Percentiles - India, Baseline



Note: Smoothing through STATA command lowess with bandwith 0.12

Figure 5: Targeting along Consumption Percentiles - Indonesia, Baseline



Note: Smoothing through STATA command lowess with bandwith 0.12

From Figure 4 we draw two conclusions. First, all models are relatively accurate at categorizing the top quintile as non-poor as well as identifying the poorest 10 percent as poor.

The misclassification rate is largely driven by households between the 10th and 60th percentile in terms of consumption. The biggest difference between perfect allocation and the predictions of the models occurs close to the rural poverty line, around the 25th percentile, where only about 40 percent of all households are classified as poor. Secondly, the paths of the different models are almost identical. The neural network, the most precise model on aggregate, seems to systematically classify more households as poor, which leads to higher inclusion and lower exclusion error rates. However, adding 95% confidence intervals to the lines of the benchmark, OLS, and the neural network shows that the models do not systematically target certain consumption percentiles differently. The respective Figure A1 can be found in the Appendix.

The graph for Indonesia, Figure 5, is slightly different, but does not alter the main conclusions. As there are lower actual poverty rates based on official poverty lines, it becomes more difficult for the models to differentiate the very poor from the non-poor. We elaborate on this in Section 6.1. As a result, already more than 20 percent of the lowest percentile are classified as non-poor, while almost all households above the 60th percentile are correctly classified. However, as Figure A2 in the Appendix confirms, OLS and the neural network do not target certain consumption percentiles differently.

We have established for the baseline of both data sets that the methods do not differ systematically in their targeting performance across the consumption distribution. However, there are more dimensions worth considering. As mentioned in Section 2.2, development programs that rely on proxy means tests to target the beneficiaries often consume significant public budgets. As a result, program directors choosing between different methods for a proxy means test, need to ensure that their choice does not discriminate along important socio-economic dimensions. For this purpose, we examine whether the methods perform differently for urban vs. rural households, for the different genders of the head of the household and for households in different states or provinces.

Table 4 and Figure 6 and 7 show the results for those three dimensions, following this regression specification:

$$\begin{aligned}\text{total error}_{imj} = \alpha + \sum_{mj=1}^{MJ} \beta_{mj} * \text{method}_{imj} * \text{dimension}_{imj} \\ + \sum_{m=1}^{3} \gamma_m * \text{method}_{imj} + \sum_{j=1}^{J} \delta_j * \text{dimension}_{imj} + \epsilon_{imj} \quad (4)\end{aligned}$$

Where the γ_m capture the method fixed effects and δ_j the fixed effects of the dimension, e.g. state or urban fixed effects. The coefficients of interests are β_{mj} which represent the difference in total error rates for the dimension of interest between the different methods.

In Table 4, we see that there are no statistically significant differences across models for urban households for India. For households whose head is female however, penalized regression and the random forest have statistically significant, but economically negligible differences in their error rates. To analyze whether one model systematically misclassifies more households in different states, we show the total error rates for all Indian states and methods in Figure 6. Visual inspection does not suggest any systematic differences across methods. For an econometric analysis, we restrict our sample to the six largest states to have sufficiently large numbers of observations to test for differences in means. We do not find any robust differences in targeting accuracy and therefore move the discussion on this to Section A.2.6 in the Appendix.

Table 4: Total Error Rates for Inspected Subgroups - Baseline

	India		Indonesia	
	(1)	(2)	(3)	(4)
Urban x PR	−0.001		0.006	
	(0.003)		(0.008)	
Urban x NN	−0.0005		0.008	
	(0.006)		(0.010)	
Urban x RF	0.002		−0.009	
	(0.005)		(0.009)	
Female x PR		0.012**		−0.003
		(0.006)		(0.011)
Female x NN		0.013		−0.007
		(0.010)		(0.007)
Female x RF		0.018**		0.011
		(0.008)		(0.010)
Constant	0.188***	0.173***	0.202***	0.157***
	(0.007)	(0.007)	(0.013)	(0.007)
Method FE	Yes	Yes	Yes	Yes
Urban FE	Yes	No	Yes	No
Female FE	No	Yes	No	Yes
Observations	41,484	41,484	11,632	11,632

Note: Standard errors are clustered at the district level.

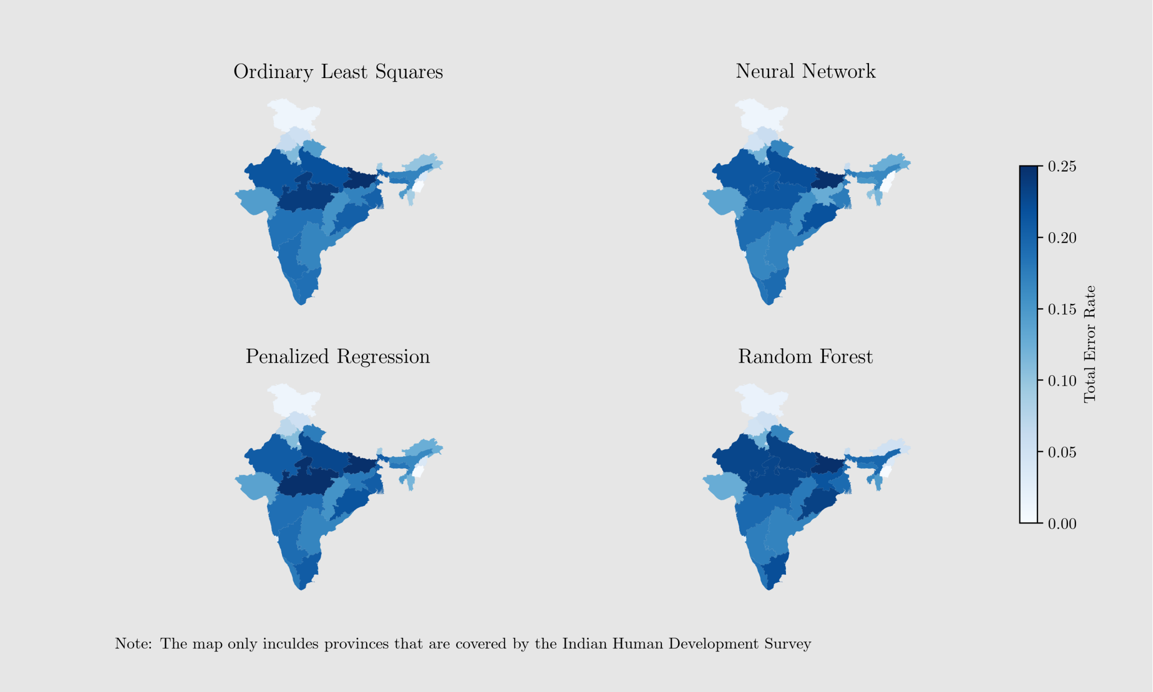
*p<0.1; **p<0.05; ***p<0.01

(1) and (3): Urban vs. Rural Households

(2) and (4): Female vs. Male Head of Household

PR: Penalized Regression, NN: Neural Network, RF: Random Forest

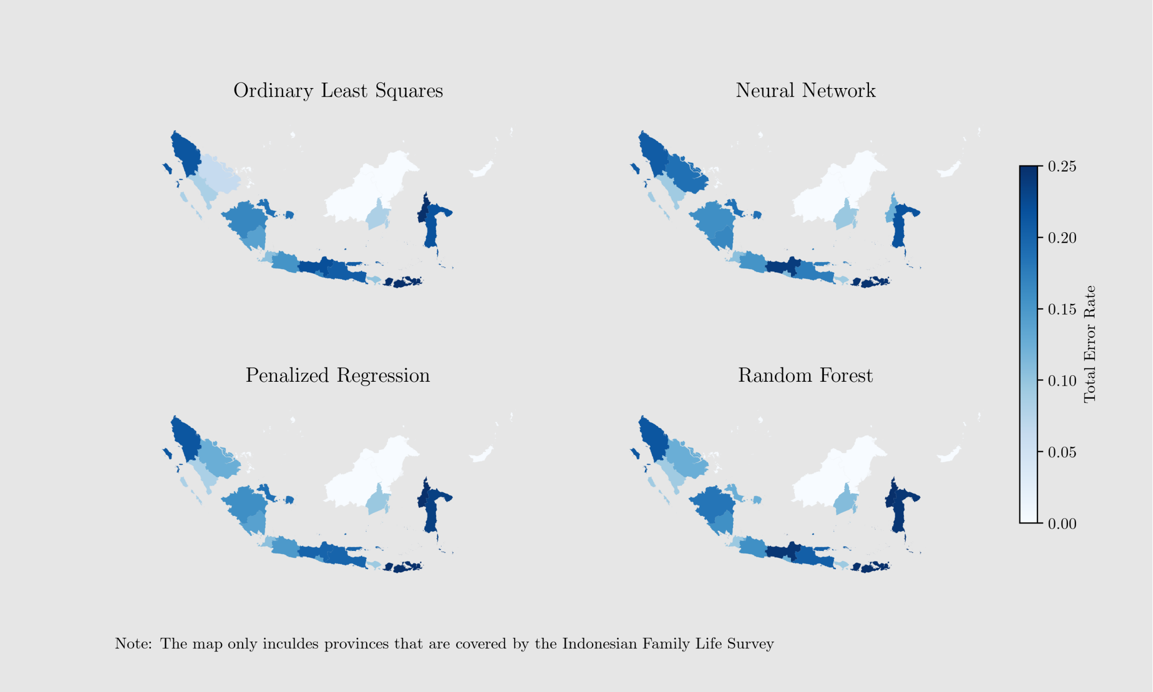
Figure 6: Total Error Rates across Indian States - Baseline Results



Note: The map only inculdes provinces that are covered by the Indian Human Development Survey

For Indonesia, we do not witness any statistically significant interactions terms for urban vs. rural households or different genders of the household head. Figure 7 depicts the total error rates across all provinces included in the data set. We observe some differences for provinces such as Riau or Lampung. However, running the same econometric analysis with the six most populated provinces shows no statistically significant differences across the methods, and we further discuss the results in Section A.2.6 of the Appendix. Overall, the small number of significant coefficients in all sub-analyses is most likely driven by the patterns found in Figures 4 and 5 which stems from the tendency of the neural networks categorizing more households as poor than OLS.

Figure 7: Total Error Rates across Indonesian Provinces - Baseline Results



Note: The map only inculdes provinces that are covered by the Indonesian Family Life Survey

5.2 Second Round Results

We repeat the analysis for the second survey round of both countries. Table 5 provides the error rates and p-values for the corresponding regressions, which are the same as equations 3 and 4.

For India, the total error rates are significantly lower across all models compared to the baseline results. This can be explained by the relatively lower official poverty level of India in 2011, which means that less households are likely to be subject to misclassification (see Section 6.1 for an intuition behind this reasoning). Regardless of this, we observe that the neural network produces the lowest total error rate, followed by the penalized regression, OLS and the random forest. However, the differences are neither economically nor statistically significant. The same results are obtained for Indonesia, where, the total error rates have also decreased compared to the baseline, but not as much as in India.

We observe a similar pattern regarding the inclusion and exclusion error rates for both data sets. OLS, penalized regressions and neural networks have very similar rates, while the random forests tend to overestimate the households' consumption compared to the other methods. This results in smaller inclusion but also higher exclusion error rates. This is mirrored in Figures A3 and A4 in the Appendix, where all methods again follow the same pattern along the consumption distribution with only the random forest categorizing households systematically as less poor.

Analogously to the first-round surveys, we examine the second-round data sets for discrimination across methods that is not explained by the small overestimation of the

Table 5: Targeting Error Rates on the Test Set - Second Round

	India								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	11.98	-	-	8.81	-	-	3.16	-	-
Penalized Regression	11.89	-0.09	0.382	9.01	0.19	0.006	2.88	-0.28	0.000
Neural Network	11.75	-0.23	0.254	9.73	0.92	0.000	2.02	-1.15	0.000
Random Forest	12.16	0.18	0.449	10.30	1.49	0.000	1.86	-1.30	0.000

Note: The test set comprises 10,370 households. Standard errors are clustered at the district level.

	Indonesia								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	14.80	-	-	10.50	-	-	4.30	-	-
Penalized Regression	14.71	-0.09	0.762	10.56	0.06	0.771	4.15	-0.14	0.416
Neural Network	14.37	-0.43	0.379	10.56	0.06	0.850	3.81	-0.48	0.183
Random Forest	15.42	0.63	0.306	12.58	2.08	0.000	2.85	-1.45	0.000

Note: The test set comprises 3,514 households. Standard errors are clustered at the district level.

random forests. As we detect no statistically and economically significant differences, we refrain from presenting the results here. Instead they can be found in Sections A.2.2 and A.2.6 of the Appendix.

6 Robustness Checks

6.1 Targeting the Extreme Poor

As mentioned in Section 2.1, poverty lines are hard to estimate precisely (Atkinson, 1987) which makes it important to assess whether our results are sensitive to the definition of the poverty line. Hence, we repeat our methodology for the first round of both countries, now calibrating all models using a poverty line that is half the level of the official ones used beforehand. This captures the extreme poor and corresponds to the 4th consumption percentile for India and the 5th consumption percentile for Indonesia.

The models now attempt to identify a very small subset of the population, located at the lower end of the consumption distribution. All methods tend to overestimate consumption at the lower end of the distribution and underestimate it at the upper end. As a result, our inclusion error becomes negligible while the exclusion error are high compared to total amount of extremely poor households. Table 6 shows the corresponding results. With the smaller target group, the differences between the methods become even smaller, leading to no significant differences, on both the aggregate and the subgroup levels. Thus, we conclude that our core findings are not sensitive to the poverty line. For completeness, we report the results of the subgroup analyses in Sections A.2.3 and A.2.6 of the Appendix.

Table 6: Targeting Error Rates on the Test Set - Half Poverty Line

	India								
	(1)			(2)			(3)		
	Total	Diff.		Excl.	Diff.		Incl.	Diff.	
	Error	to OLS	p-value	Error	to OLS	p-value	Error	to OLS	p-value
Ord. Least Squares	3.78	-	-	3.29	-	-	0.49	-	-
Penalized Regression	3.76	-0.02	0.725	3.38	0.10	0.015	0.38	-0.12	0.001
Neural Network	3.64	-0.13	0.200	2.90	-0.39	0.000	0.74	0.25	0.001
Random Forest	3.68	-0.10	0.351	3.43	0.14	0.118	0.25	-0.24	0.001

Note: The test set comprises 10,371 households. Standard errors are clustered at the district level.

	Indonesia								
	(1)			(2)			(3)		
	Total	Diff.		Excl.	Diff.		Incl.	Diff.	
	Error	to OLS	p-value	Error	to OLS	p-value	Error	to OLS	p-value
Ord. Least Squares	4.61	-	-	4.30	-	-	0.31	-	-
Penalized Regression	4.61	-0.00	1.000	4.26	-0.03	0.708	0.34	0.03	0.740
Neural Network	4.71	0.10	0.623	3.75	-0.55	0.000	0.96	0.65	0.001
Random Forest	4.81	0.21	0.222	4.78	0.48	0.020	0.03	-0.28	0.002

Note: The test set comprises 2,908 households. Standard errors are clustered at the district level.

6.2 Using a Short Vector of Input Variables

So far, we have simulated a setting where around 40 characteristics are used to calibrate the models and predict consumption. However, in practice, program directors are often restricted to shorter surveys as the information must be collected for the whole population covered by the program. For example, in India, the nationwide census from 2011 was comprised of only 29 questions in total (Ministry of Home Affairs, Government of India, 2011) and McBride and Nichols (2016) also only use around 20 characteristics for their comparison of econometric methods and random forests. Shorter questionnaires also lead to better data quality and less misreporting (Niehaus et al., 2013). Additionally, households adjust their behavior to become eligible for a given social program (Glewwe et al., 1989; Martinelli and Parker, 2009). They can, for example, hide their TV when visited by the interviewers if TV ownership is a determinant for eligibility of said program. Hence, we create short vectors for the first survey round of both countries to assess whether the methods perform differently from each other when using a smaller set of input variables.

To select the variables for the short vector, we follow Varian (2014) and utilize a quantitative ranking of the predictive power of the input variables produced by the random forest. A variable's importance is determined by computing the average decrease in prediction errors resulting from the introduction of that variable into the forest (Friedman et al., 2001). We obtain a list of the 25 variables with the highest predictive power for consumption, according to the random forest. Following the findings of Martinelli and Parker (2009), we drop variables that are easy to hide for households. For example, for Indonesia, we exclude the dummy variables that indicate ownership of a fridge or a TV. Table 7 lists the input variables for the short vector of both data sets.

Table 7: Overview of Variables Used for the Short Vector

Country	Data Set	Selected Variables	Count
India	IHDS1	# of persons, persons_sq, education hh head, size loans, dep. ratio, education oldest besides hh head, age hh head, age_sq, number rooms, gov relations, solid floor, # of kids, # of kids in school, ownership of house, state, region	16
Indonesia	IFLS4	# of persons, persons_sq, hh size per capita, education level hh head, type cooking fuel, type ownership of house, age hh head, age_sq, dep. ratio, distance to clinic, distance to posyandu, # of kids in school, education level oldest male, province, region	15

Table 8 provides the error rates for the two short sets. As expected, we observe higher total error rates compared to Section 5.1 because the methods have less information available. Interestingly, however, the error rates in both countries remain below 20 percent for all models which suggests that the marginal gain in adding more input variables is small.

For India, OLS, penalized regression and random forest have similar error rates, the model of the neural network again yields the lowest error rate. However, despite being statistically significant, the difference is only about 0.8 percentage points. In the case of Indonesia, the total error rates of the models do not differ significantly. The neural network yields the lowest error rate, followed by the penalized regression, OLS and the random forest. In both countries, the neural network tends to predict slightly lower consumption levels. As a result, it has higher inclusion error rates of up to 1.6 percentage points compared to OLS but even lower exclusion error rates. Analogously to the previous chapter, we analyze whether there are systematic differences across methods for the models calibrated on the small set of input variables. As we do not detect such differences, we report all analyses in Sections A.2.4 and A.2.6 of the Appendix.

Table 8: Targeting Error Rates on the Test Set - Short Vector

	India								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	19.39	-	-	12.97	-	-	6.42	-	-
Penalized Regression	19.60	0.21	0.118	13.40	0.43	0.000	6.20	-0.22	0.027
Neural Network	18.55	-0.84	0.009	10.54	-2.43	0.000	8.01	1.59	0.000
Random Forest	19.28	-0.11	0.668	13.60	0.63	0.004	5.69	-0.73	0.001

Note: The test set comprises 10,371 households. Standard errors are clustered at the district level.

	Indonesia								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	18.12	-	-	14.20	-	-	3.92	-	-
Penalized Regression	17.98	-0.14	0.290	14.27	0.07	0.492	3.71	-0.21	0.017
Neural Network	17.54	-0.58	0.092	12.79	-1.41	0.000	4.75	0.83	0.000
Random Forest	18.33	0.21	0.595	14.51	0.31	0.429	3.82	-0.10	0.781

Note: The test set comprises 2,908 households. Standard errors are clustered at the district level.

7 Model Stability over Time

As mentioned in Section 2.1, surveys containing consumption data are expensive and can therefore not be conducted every year. For example, the two survey rounds for India lie 6 years apart and the surveys for Indonesia are carried out every 7 years. Consequently, program directors might not be able to re-calibrate their models regularly, forcing them to rely on the old models when new households need to be assessed. This can become a problem in countries such as India or Indonesia, where societies have been undergoing a rapid development process and the underlying relationships between household characteristics and poverty could have changed. For example, between the two survey rounds in Indonesia, the share of households in the survey that uses gas stoves for cooking went up from 18 percent in 2007 to 70 percent in 2014, largely due to the government's Kerosene to Liquefied Petroleum Gas program (Desai et al., 2010; 2015; Zhang, 2013). A model that had assigned a significant impact to the characteristic "main fuel used for cooking" in 2007 would consequently misclassify many households in 2013.

Hence, we analyze whether some methods cope better with these structural changes. To answer this question, we compare the prediction performance of models that have been calibrated based on the first survey round data and evaluate their consumption predictions for the subsequent survey round. For this purpose, we slightly adjust our training and evaluation processes compared to Section 5 and 6. We now train all models on the full data set from the first survey but keep the 67/33 split for the hyperparameter tuning for penalized regressions, neural networks and random forests. However, before we optimize the models, we need to consider that prices have risen between survey rounds. As our consumption variables are measured in nominal terms, we inflate the first-round values using official inflation rates from the World Bank (2019). We consider this a sensible approach, as inflation data would be available for program directors as well. Finally, we optimize the methods as described in Section 4 and evaluate their prediction performance using the second round data as the test set.

Table 9: Targeting Error Rates on the Test Set - Time Stability

	India								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	12.37	-	-	8.02	-	-	4.35	-	-
Penalized Regression	12.50	0.13	0.012	8.09	0.07	0.062	4.41	0.06	0.134
Neural Network	12.50	0.13	0.403	8.72	0.70	0.000	3.78	-0.58	0.000
Random Forest	13.15	0.78	0.000	7.16	-0.86	0.000	6.00	1.65	0.000

Note: The test set comprises 41,491 households. Standard errors are clustered at the district level.

	Indonesia								
	(1)			(2)			(3)		
	Total Error	Diff. to OLS	p-value	Excl. Error	Diff. to OLS	p-value	Incl. Error	Diff. to OLS	p-value
Ord. Least Squares	15.75	-	-	11.65	-	-	4.10	-	-
Penalized Regression	15.92	0.17	0.372	11.63	-0.02	0.890	4.29	0.19	0.163
Neural Network	16.13	0.38	0.239	13.40	1.75	0.000	2.72	-1.37	0.000
Random Forest	16.03	0.28	0.352	12.28	0.63	0.000	3.75	-0.35	0.274

Note: The test set comprises 14,056 households. Standard errors are clustered at the district level.

The results for the time stability analysis are displayed in Table 9. As expected, the error rates are higher in this analysis compared to Section 5.2, where the models are both trained and evaluated on the second-round surveys. Surprising is rather, that the difference is less than 2 percentage points across all models although more than five years lie between the surveys. This suggests that not only poverty is chronic, but also the characteristics that predict poverty do not fundamentally change over time.

For India, the exclusion errors have slightly decreased, while the inclusion errors have increased. A reason for this might be that the average consumption level in the first survey round after adjusting for inflation is still lower than the one in the second survey round. This means that we tend to underestimate consumption and thus include more ineligible households in the program. For Indonesia, we find no clear pattern when comparing inclusion and exclusion error rates to the second round. Overall, we do not observe large differences in aggregate error rates when comparing the targeting of the models against each other. OLS has the lowest total error rate of all the models in both data sets, but in all cases besides the random forest model for India, the error rates of the other models do not differ significantly from each other. This suggests that the methods do not differ strongly in their ability to capture patterns between consumption and household characteristics over time. This is confirmed by the respective subgroup results, reported in Sections A.2.5 and A.2.6 of the Appendix.

8 Discussion

Several discussion points arise from our results. First of all, we notice that no method performs consistently better in all analyses. Neural networks seem to be slightly more precise in many cases, penalized regressions and random forests are comparable in precision to OLS. The differences between all methods are often not statistically significant and always economically small. Nevertheless, little improvements in aggregate misclassification rates can have a big impact for individual families as receiving a legal entitlement for participating in an anti-poverty program can be crucial for concerned households. Hence, if program directors have sufficient time and resources at hand, they might try out different methods in an out-of-sample validation process to find out which method works best for a given setting, as has been done in this thesis. However, in practice, time and budget constraints exist and most likely prevail.

Nonetheless, we recommend to use random forests to select the best variables to include in the household surveys. As pointed out before, only short questionnaires are used in practice when it comes to surveying the entire population. We have shown that reducing the number of questions from 42 for the Indian survey and 35 in the Indonesian survey to 16 and 15 respectively increases the error rates only slightly. Thus, we consider the variable selection process of the random forest as effective and applicable in practice. This would also allow researchers to explore data sets with even larger number of variables effectively.

Overall however, it seems clear to us that, given the current restrictions on data availability for proxy means tests, no complex statistical method is able to reduce the substantial misclassification rates to negligible numbers. Under this impression, it seems more promising to explore approaches that do not solely rely on these methods for targeting. For example, Alatas et al. (2016) investigate a combination of proxy means test and self-targeting mechanisms in Indonesia. The authors show that introducing a small but significant application cost before conducting the proxy means test can lead to better targeting outcomes than pure proxy means tests. This is because non-poor households do not even apply as they are unlikely to become eligible. Similar mechanisms could also be used in combination with universal entitlements which theoretically ensure that no households are excluded (Slater, 2011).

Our selection of variables might actually be a limiting factor, as we stick close to the literature and use the same variables for all models in order to get a strong and representative OLS benchmark. Yet, given one of the main advantages of the machine learning tools is to capture non-linear functional forms and interactions, we speculate that there are not enough sensible patterns and combinations of variables that can be distilled from the data sets. It might have been fruitful to consider other variables that researchers have not yet tried out as well. Here, the field might benefit from continued research and the evaluation of the methods against each other for different types of variables. This would need to be combined with research on survey design. For instance, researchers might try to integrate more questions like "Do you have access to a public tap?". This could be relevant and hence good for predicting consumption for poor households, as a public tap represents a cheap and healthy source of water. For households in the upper income percentiles however, this will most likely be irrelevant as they have private taps installed in their homes.

Additionally, we want to emphasize the ambiguity of using machine learning tools for predictions in poverty targeting. As pointed out by several authors (Cameron and Shah, 2013; Lavallée et al., 2010), one disadvantage of using proxy means tests for poverty targeting is that regression analysis is difficult to explain. We believe that this problem would be aggravated in the case of using, for instance, a neural network as its predictions are even harder to explain than those generated by OLS. Consequently, the perceived fairness and acceptance of such a method might be even lower. However, not having a simple linear scoring model also bears a significant advantage. It makes it harder for locals to hide assets or underreport other characteristics that can be identified as the most important ones for the classification decision (Camacho and Conover, 2011). As recommended by Blumenstock (2018), further applied research is needed to assess whether the advantages outweigh the disadvantages when applying machine learning to poverty targeting.

Another limitation is the definition of the metrics we used for calculating the misclassification rates, which we kept simple to make them easily accessible and interpretable. Yet, when doing the robustness check with the reduced poverty line, we witness the shortcomings of our metrics. The misclassification rates plummet, making it hard to compare results across countries with different poverty lines. The drop could wrongly be interpreted as an improved targeting performance but is actually due to the lower poverty line, as discussed in Section 6.1. More complex evaluation metrics such as undercoverage and leakage rates used by McBride and Nichols (2016) can overcome these limitations. Also, our approach to weigh inclusion error and exclusion equally to compute the total error rate might be subject to criticism. As poverty lines themselves are controversial, other authors have tried using poverty rates for their targeting approaches instead (Brown et al., 2016). However, this discussion is not critical for the specific purpose of our thesis, as our metrics still allow a valid comparison of the four methods.

Furthermore, we do not discuss poverty outcomes in our thesis. Assessing the actual impact on poverty of the slight targeting improvements we found would be helpful to make a better cost-benefit assessment of machine learning tools in poverty targeting.

Given our results and the limitations above, we can discuss the external validity of this thesis. With India and Indonesia, our analysis is based on data from two countries in South and Southeast Asia. Hence, our results could differ in other developing countries, for example, in South America or Sub-Saharan Africa. However, the data restrictions regarding the number of variables used and the sample size of households available for calibration are likely to be similar in other settings. As witnessed in McBride and Nichols (2016), the sample sizes in practice range rather between 1,500 and 12,000 observations than around 40,000 as in the India data set we use. The work by McBride and Nichols (2016) can also be seen as an indicator of how well our analysis could translate to other settings, as they analyze three countries from South America, Sub-Saharan Africa and Southeast Asia and find error rates comparable to ours. Therefore, we do not have any indication that our research question would be answered differently for proxy means tests in other developing countries.

However, other researchers have shown that machine learning tools can be successfully applied to important problems in development economics. Jean et al. (2016) use high-resolution satellite images and nighttime light intensity data on over 300,000 locations and convolutional neural networks to predict poverty accurately on a local level, where on average 30 households are clustered to a local unit. Another example is Blumenstock et al. (2015), who use mobile calls and transfer data on more than 1,000,000 unique users in Rwanda to shed light on consumption smoothing mechanisms in the aftermath of natural disasters. This suggests that using larger data sets for PMTs might yield different results as machine learning tools are expected to extract information from large data sets particularly well (Varian, 2014).

Overall, machine learning tools can be successfully applied for out-of-sample prediction tasks in data-rich settings in business and economics (Desai et al., 1996; Einav and Levin, 2014; Mullainathan and Spiess, 2017). However, as the *no free lunch* theorem states, no learning and prediction tool is best suited for every problem which creates the need for further empirical research (Wolpert, 1996). Apart from prediction tasks, there is an increasing amount of research on the application of machine learning tools to estimate causal effects, for example, in the context of instrument variables (Hartford et al., 2017) or synthetic control methods (Athey and Imbens, 2017). Nevertheless, as Blumenstock (2018) discusses, machine learning tools not only need to fulfill certain data requirements but also need to be applied carefully to overcome their pitfalls, such as bias in the algorithm or lack of regulation and transparency.

9 Conclusion

In this thesis, we compare the out-of-sample prediction performance of different statistical methods in the context of proxy means tests. Specifically, we assess whether penalized regressions, neural networks or random forests can be more accurate than ordinary least squares in identifying which households become eligible for anti-poverty programs.

To do so, we use two multi-topic, panel household surveys for India and Indonesia which we randomly split into training and test sets. The former are used to develop the optimal models which, in turn, predict the consumption levels of the households in the test sets. This way, we can assess the out-of-sample prediction performance of the different methods. For this purpose, we compare their total, exclusion and inclusion error rates using official national poverty lines. Additionally, we investigate whether the methods target important subgroups differently, such as certain consumption percentiles, rural households, households whose head is female or households in certain states/provinces. Finally, we analyze if any method offers significant advantages when predicting consumption using training data from previous periods.

Overall, we find that there is little difference in the prediction accuracy of the four methods. Neural networks yield the lowest total error rates in eight of our ten analyses. However, the difference is statistically significant only for two out of these settings and even then, the differences are small in an economic sense. Our analyses reveal no systematic differences across methods when it comes to targeting important socio-economic or geographic subgroups. This pattern is robust to settings with a short set of variables, a poverty line at half the official level and holds over time. Surprisingly, all methods do very well in predicting consumption for the second survey round, even though the data they are trained with was collected more than five years earlier.

We draw one important policy recommendation from our results. We propose using random forests to optimize the surveys on which program eligibility is determined in targeting programs. Our results in Section 6.2 suggest that they are well-suited to identify the variables with the highest prediction power as the total error rate drops by 2.1 percentage points at most despite only using half the number of characteristics. As the length of a survey is an important determinant of the cost of collection and the quality of data, this tool can be helpful in improving a program's efficiency.

Our research suffers from several limitations. We explore the methods only for data sets of two countries from South and Southeast Asia and thus cannot ensure that our results will hold for other developing countries. Also, we restrict our comparison to four methods and keep our study concise by defining only simple outcome metrics. Future research could also take full advantage of the machine learning methods by including other non-traditional variables connected with consumption through non-linear relationships. Consequently, it might be fruitful to optimize the surveys accordingly and assess whether further improvements are possible. And finally, it is crucial to consider the impact of machine learning methods on transparency and misreporting concerns. Hence, future research should also discuss the costs and benefits of having models that might be more precise but are also harder to communicate to the public compared to traditional methods.

Machine learning techniques have become powerful tools with significant advantages over traditional econometric methods in data-rich settings. Studies like the ones by Blumenstock et al. (2016) or Jean et al. (2016) have shown that there are also successful applications of machine learning in development economics, for instance, for estimating poverty levels on a local level or explaining consumption patterns. Unfortunately, proxy means tests currently do not represent such data-rich settings in practice and the application of machine learning methods can consequently only lead to marginal improvements.

References

Ahrens, A., Hansen, C. and Schaffer, M. (2018), 'cvlasso: Program for cross-validation using lasso, square-root lasso, elastic net, adaptive lasso and post-ols estimators'.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A. and Tobias, J. (2012), 'Targeting the poor: evidence from a field experiment in indonesia', *American Economic Review* **102**(4), 1206–40.

Alatas, V., Purnamasari, R., Wai-Poi, M., Banerjee, A., Olken, B. A. and Hanna, R. (2016), 'Self-targeting: Evidence from a field experiment in indonesia', *Journal of Political Economy* **124**(2), 371–427.

Asra, A. (2000), 'Poverty and inequality in indonesia: estimates, decomposition and key issues', *Journal of the Asia Pacific Economy* **5**(1-2), 91–111.

Athey, S. and Imbens, G. W. (2017), 'The state of applied econometrics: Causality and policy evaluation', *Journal of Economic Perspectives* **31**(2), 3–32.

Atkinson, A. B. (1987), 'On the measurement of poverty', *Econometrica: Journal of the Econometric Society* pp. 749–764.

Atkinson, A. B. and Bourguignon, F. (1982), 'The comparison of multi-dimensioned distributions of economic status', *The Review of Economic Studies* **49**(2), 183–201.

Bennett, J. G. (2017), 'Poverty targeting primer. concepts, methods and tools'.

Berg, J. (2010), 'Conditional cash transfers as response to the crisis - the bolsa familia programme'.

Bergstra, J. and Bengio, Y. (2012), 'Random search for hyper-parameter optimization', *Journal of Machine Learning Research* **13**(Feb), 281–305.

Biswas, S. (2019), 'India election: Why rahul gandhi's minimum income plan is a gamble, bbc'. Accessed: 2019-05-09.
URL: *https://www.bbc.com/news/world-asia-india-47038421*

Blumenstock, J. (2018), 'Don't forget people in the use of big data for development'.

Blumenstock, J., Cadamuro, G. and On, R. (2015), 'Predicting poverty and wealth from mobile phone metadata', *Science* **350**(6264), 1073–1076.

Blumenstock, J. E., Eagle, N. and Fafchamps, M. (2016), 'Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters', *Journal of Development Economics* **120**, 157–181.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Brown, C., Ravallion, M. and Van de Walle, D. (2016), *A poor means test? Econometric targeting in Africa*, The World Bank.

Camacho, A. and Conover, E. (2011), 'Manipulation of social program eligibility', *American Economic Journal: Economic Policy* **3**(2), 41–65.

Cameron, L. and Shah, M. (2013), ‘Can mistargeting destroy social capital and stimulate crime? evidence from a cash transfer program in indonesia’, *Economic Development and Cultural Change* **62**(2), 381–415.

Chollet, F. (2018), *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*, MITP-Verlags GmbH & Co. KG.

Chollet, F. et al. (2015), ‘Keras’, https://keras.io. Accessed: 2019-05-09.

Coady, D., Grosh, M. and Hoddinott, J. (2004*a*), *Targeting of transfers in developing countries: Review of lessons and experience*, The World Bank.

Coady, D., Grosh, M. and Hoddinott, J. (2004*b*), ‘Targeting outcomes redux’, *The World Bank Research Observer* **19**(1), 61–85.

Daly, A. and Fane, G. (2002), ‘Anti-poverty programs in indonesia’, *Bulletin of Indonesian Economic Studies* **38**(3), 309–329.

Deaton, A. (1997), *The analysis of household surveys: a microeconometric approach to development policy*, The World Bank.

Deaton, A. and Dreze, J. (2002), ‘Poverty and inequality in india: a re-examination’, *Economic and political weekly* pp. 3729–3748.

Deaton, A. and Zaidi, S. (2002), *Guidelines for constructing consumption aggregates for welfare analysis*, Vol. 135, World Bank Publications.

Desai, S., Vanneman, R. and National Council of Applied Economic Research, N. D. (2010), ‘India human development survey (ihds), 2005. icpsr22626-v8’.

Desai, S., Vanneman, R. and National Council of Applied Economic Research, N. D. (2015), ‘India human development survey-ii (ihds-ii), 2011-12. icpsr36151-v2’.

Desai, V. S., Crook, J. N. and Overstreet Jr, G. A. (1996), ‘A comparison of neural networks and linear scoring models in the credit union environment’, *European Journal of Operational Research* **95**(1), 24–37.

Duflo, E. (2001), ‘Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment’, *American economic review* **91**(4), 795–813.

Dutrey, A. P. (2007), *Successful targeting?: reporting efficiency and costs in targeted poverty alleviation programmes*, United Nations Research Institute for Social Development Geneva, Switzerland.

Einav, L. and Levin, J. (2014), ‘The data revolution and economic analysis’, *Innovation Policy and the Economy* **14**(1), 1–24.

Friedman, J., Hastie, T. and Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Gertler, P. (2004), ‘Do conditional cash transfers improve child health? evidence from progresa’s control randomized experiment’, *American economic review* **94**(2), 336–341.

Gillis, M., Shoup, C. and Sicat, G. P. (2001), *World development report 2000/2001-attacking poverty*, The World Bank.

Glewwe, P., Kanaan, O. et al. (1989), *Targeting assistance to the poor using household survey data*, Vol. 225, World Bank.

Government of India Planning Commission (2009), 'Report of the expert group to review the methodology for estimation of poverty'. Accessed: 2019-05-09.
URL: *http://planningcommission.nic.in/reports/genrep/rep_pov.pdf*

Government of India Planning Commission (2013), 'Press note on poverty'. Accessed: 2019-05-09.
URL: *http://planningcommission.nic.in/news/pre_pov2307.pdf*

Grosh, M. E. and Baker, J. L. (1995), *Proxy means tests for targeting social programs: simulations and speculation*, The World Bank.

Grosh, M. and Glewwe, P. (2000), 'Designing household questionnaires for developing countries: lessons from 15 years of the livings standards measurement study'.

Hanna, R. and Olken, B. A. (2018), 'Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries', *Journal of Economic Perspectives* **32**(4), 201–26.

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2017), Deep iv: A flexible approach for counterfactual prediction, *in* 'Proceedings of the 34th International Conference on Machine Learning-Volume 70', JMLR. org, pp. 1414–1423.

Haykin, S. (1994), *Neural networks: a comprehensive foundation*, Prentice Hall PTR.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016), 'Combining satellite imagery and machine learning to predict poverty', *Science* **353**(6301), 790–794.

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Ijcai', Vol. 14, Montreal, Canada, pp. 1137–1145.

Lavallée, E., Olivier, A., Pasquier-Doumer, L., Robilliard, A.-S. et al. (2010), 'Poverty alleviation policy targeting: a review of experiences in developing countries', *Document de Travail. Paris: Université Paris-Dauphine/IRD* .

LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *nature* **521**(7553), 436.

Litvack, J. I. (2011), 'Social safety nets: An evaluation of world bank support, 2000–2010'.

Martinelli, C. and Parker, S. W. (2009), 'Deception and misreporting in a social program', *Journal of the European Economic Association* **7**(4), 886–908.

McBride, L. and Nichols, A. (2016), 'Retooling poverty targeting using out-of-sample validation and machine learning', *The World Bank Economic Review* **32**(3), 531–550.

Ministry of Home Affairs, Government of India (2011), 'Census of india 2011 - household schedule'. `http://censusindia.gov.in/2011-Schedule/Shedules/English_Household_schedule.pdf, Accessed: 2019-05-09`.

Mkandawire, T. (2005), *Targeting and universalism in poverty reduction*, United Nations Research Institute for Social Development Geneva.

Mullainathan, S. and Spiess, J. (2017), 'Machine learning: an applied econometric approach', *Journal of Economic Perspectives* **31**(2), 87–106.

Narayan, A. and Yoshida, N. (2005), 'Proxy means tests for targeting welfare benefits in sri lanka', *Report No. SASPR–7, Washington, DC: World Bank* **5**, 2009.

Niehaus, P., Atanassova, A., Bertrand, M. and Mullainathan, S. (2013), 'Targeting with agents', *American Economic Journal: Economic Policy* **5**(1), 206–38.

Orshansky, M. (1963), 'Children of the poor', *Soc. Sec. Bull.* **26**, 3.

Papageorgiou, G., Grant, S. W., Takkenberg, J. J. M. and Mokhles, M. M. (2018), 'Statistical primer: how to deal with missing data in scientific research?†', *Interactive CardioVascular and Thoracic Surgery* **27**(2), 153–158.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine Learning in Python ', *Journal of Machine Learning Research* **12**, 2825–2830.

Priebe, J. (2014), 'Official poverty measurement in indonesia since 1984: a methodological review', *Bulletin of Indonesian Economic Studies* **50**(2), 185–205.

Schultz, T. P. (2004), 'School subsidies for the poor: evaluating the mexican progresa poverty program', *Journal of development Economics* **74**(1), 199–250.

Sen, A. (1973), 'Poverty, inequality and unemployment: some conceptual issues in measurement', *Economic and Political Weekly* pp. 1457–1464.

Slater, R. (2011), 'Cash transfers, social protection and poverty reduction', *International Journal of Social Welfare* **20**(3), 250–259.

Strauss, J., Witoelar, F. and Sikoki, B. (2016), 'The fifth wave of the indonesia family life survey (ifls5): Overview and field report'.

Strauss, J., Witoelar, F., Sikoki, B. and Wattie, A. (2009), 'The fourth wave of the indonesia family life survey (ifls4): Overview and field report'.

Subbarao, K. and Smith, W. J. (2003), 'What role for safety net transfers in very low income countries?', *Paper, Social Safety Net Primer Series, World Bank, Washington DC* .

Subramanian, S. and Deaton, A. (1996), 'The demand for food and calories', *Journal of political economy* **104**(1), 133–162.

The World Bank (2019), 'World development indicators'.

Tu, J. V. (1996), 'Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes', *Journal of clinical epidemiology* **49**(11), 1225–1231.

Varian, H. R. (2014), ‘Big data: New tricks for econometrics’, *Journal of Economic Perspectives* **28**(2), 3–28.

Wolpert, D. H. (1996), ‘The lack of a priori distinctions between learning algorithms’, *Neural computation* **8**(7), 1341–1390.

World Bank Group (2018), ‘Brazil boost public expenditure database’.

Zhang, Yabei; Tuntivate, V. A. C. W. Y. (2013), ‘Indonesia - toward universal access to clean cooking’, *East Asia and Pacific (EAP) clean stove initiative knowledge exchange series* .

Zou, H. and Hastie, T. (2005), ‘Regularization and variable selection via the elastic net’, *Journal of the royal statistical society: series B (statistical methodology)* **67**(2), 301–320.

A Appendix

A.1 Data and Methodology

In this Subsection of the Appendix, we present all materials that serve as supplements to better understand the data used and methodology applied.

A.1.1 Descriptive Statistics

We split both the Indian and the Indonesian data set into training and test sets, as explained in Section 4. To make sure this randomized split has worked and we do not have different population groups in the two parts, we do provide tables to compare the means of the training and test sets for the first round surveys. Table A1 refers to the Indian data set. Column (1) depicts the mean of the training set and Column (2) the mean of the test set. Column (3) shows the difference in means and Column (4) the p-value of a t-test for difference in means. We can infer that all variables are not significantly different between the training and test set. The same is done for the Indonesian data set in Table A2, with the same result that the training and test set are balanced.

Table A1: Balance of Training and Test Set - India

	India			
Variable	Training	Test	Diff.	p-Value
cons	957.78	947.00	10.78	0.353
urb	0.36	0.35	0.01	0.288
edu oldest	7.56	7.55	0.02	0.79
nfarm	0.75	0.78	-0.03	0.087
person	5.2	5.19	0.00	0.877
cow	0.29	0.29	0.00	0.97
buffalo	0.26	0.25	0.00	0.706
bike	0.55	0.55	0.00	0.906
motorbike	0.19	0.19	0.00	0.784
fan	0.64	0.64	0.00	0.932
telephone	0.17	0.17	0.00	0.776
cell	0.09	0.09	0.00	0.618
fridge	0.18	0.18	0.00	0.863
size loan	18,920.60	18,531.00	389.60	0.685
size	2.59	2.59	0.00	0.799
elec	0.78	0.78	-0.01	0.263
incfarm	8,372.90	7,410.00	962.90	0.075
tv	0.54	0.54	0.00	0.759
d water2	0.46	0.46	0.00	0.834
d water3	0.24	0.24	0.00	0.56
d water4	0.27	0.27	0.00	0.459
d water5	0.01	0.01	0.00	0.235
own2	0.09	0.09	0.00	0.376
own3	0.02	0.02	0.00	0.091
floor	0.58	0.58	0.00	0.678
wall	0.62	0.63	0.00	0.609
toilet	0.46	0.46	0.00	0.996
roof	0.51	0.51	0.00	0.878
wood	0.68	0.68	0.00	0.88
person sq	33.12	33.33	-0.22	0.629
gov hh	0.35	0.34	0.00	0.536
medical	0.48	0.47	0.00	0.687
kitchen2	0.61	0.61	0.00	0.845
kitchen3	0.2	0.2	0.00	0.429
caste2	0.17	0.17	0.00	0.954
caste3	0.34	0.34	0.00	0.414
caste4	0.2	0.2	0.00	0.657
caste5	0.08	0.08	0.00	0.532
caste6	0.11	0.11	0.00	0.619
caste7	0.02	0.02	0.00	0.963
caste8	0.02	0.02	0.00	0.47
sex	1.1	1.1	0.00	0.706
age	47.15	46.91	0.24	0.123
edu head	5.52	5.5	0.02	0.775
children	0.47	0.48	-0.01	0.345
dratio	0.74	0.73	0.00	0.672
nschool	1.35	1.34	0.02	0.312
disabled	0.11	0.11	0.00	0.936
age sq	2,404.13	2,381.00	23.13	0.134
widow	0.1	0.1	0.00	0.805
married	0.87	0.87	0.00	0.957
d occ2	0.04	0.04	0.00	0.347
d occ3	0.01	0.01	0.00	0.896
d occ4	0.05	0.05	0.00	0.906
d occ5	0.02	0.01	0.00	0.195
d occ6	0.04	0.04	0.00	0.304
d occ7	0.18	0.19	-0.01	0.066
d occ8	0.12	0.12	0.00	0.524
d occ9	0.06	0.06	0.00	0.703
d occ10	0.05	0.05	0.00	0.779
Observations	31,117	10,371	41,484	41,484

Table A2: Balance of Training and Test Set - Indonesia

	Indonesia			
Variable	Training	Test	Diff.	p-Value
cons	1,298,908.00	817,418.00	481,490.00	0.214
floor	0.52	0.51	0.01	0.22
wall	0.73	0.73	0.00	0.945
urb	0.54	0.53	0.01	0.32
elec	0.96	0.96	0.00	0.254
borrow	0.88	0.89	-0.01	0.058
person	5.34	5.28	0.06	0.346
children	0.41	0.42	-0.01	0.617
person sq	37.47	37.22	0.25	0.804
dratio	1.36	1.38	-0.02	0.467
age	43.52	43.36	0.16	0.633
sex	0.81	0.82	-0.01	0.424
age sq	2,122.73	2,108.00	14.73	0.636
farm dummy	0.28	0.27	0.01	0.302
farm size	4,990.20	4,235.00	755.20	0.759
clinic distance	18.25	18.24	0.01	0.996
posyandu distance	9.31	8.85	0.46	0.577
clinic knowledge	0.89	0.89	0.00	0.96
posyandu knowledge	0.78	0.78	0.00	0.977
size per capita	22.97	19.18	3.79	0.3
toilet type2	0.09	0.1	-0.01	0.103
toilet type3	0.14	0.13	0.01	0.47
toilet type4	0.02	0.02	0.00	0.458
toilet type5	0.00	0.00	0.00	0.864
water type2	0.51	0.51	0.01	0.611
water type3	0.09	0.09	0.00	0.88
water type4	0.01	0.01	0.00	0.037
water type5	0.17	0.16	0.01	0.34
water type6	0.00	0.00	0.00	0.748
cook type2	0.18	0.17	0.01	0.131
cook type3	0.41	0.42	-0.01	0.346
cook type4	0.35	0.36	-0.01	0.602
cook type5	0.04	0.04	0.00	0.81
own type2	0.19	0.2	-0.01	0.408
own type3	0.11	0.12	0.00	0.961
marstat type2	0.78	0.8	-0.01	0.248
marstat type3	0.01	0.01	0.00	0.558
marstat type4	0.02	0.02	0.00	0.776
marstat type5	0.10	0.10	0.01	0.305
occ type2	0.28	0.28	0.00	0.674
occ type3	0.01	0.01	0.00	0.613
occ type4	0.10	0.10	.000	0.594
occ type5	0.00	0.00	0.00	0.116
occ type6	0.05	0.05	0.00	0.831
occ type7	0.17	0.17	0.00	0.918
occ type8	0.04	0.04	0.01	0.144
occ type9	0.01	0.01	0.00	0.567
occ type10	0.17	0.18	-0.01	0.385
occ type11	0.00	0.00	0.00	0.564
work type2	0.43	0.42	0.01	0.228
work type3	0.07	0.07	-0.01	0.12
work type4	0.25	0.24	0.00	0.625
work type5	0.1	0.1	0.00	0.555
edulev head2	0.4	0.38	0.02	0.068
edulev head3	0.15	0.16	-0.01	0.231
edulev head4	0.35	0.36	0.00	0.879
nschool	0.69	0.67	0.02	0.384
edulev hhm2	0.15	0.14	0.01	0.079
edulev hhm3	0.21	0.22	-0.01	0.217
edulev hhm4	0.63	0.63	0.00	0.889
dratiodum	0.95	0.96	0.00	0.643
tvdum	0.74	0.73	0.01	0.233
fridgeown	0.40	0.39	0.01	0.522
fridgeused	0.28	0.27	0.00	0.789
Observations	8,723	2,908	11,631	11,631

A.1.2 Cross Validation

Cross validation as a concept for accuracy estimation and model selection in statistics has already been explored in the 1990s (Kohavi et al., 1995). In recent years, it has become a popular tool for optimizing hyperparameters in machine learning (Chollet, 2018). In this particular context, it is often referred to as k-fold cross validation. For an intuitive explanation, we heavily rely on Varian (2014), who uses five steps to explain the algorithm. Suppose, we want to find the optimal value for the parameter α of model m on a given data set D. Then the k-fold cross validation will work as follows:

1. The algorithm splits the data into k equal subsets, called folds. They are labeled $s = 1, ..., k$, where $s = 1$ is the starting value.

2. A value for the parameter α is chosen (consider that for picking this value, we might have different strategies, such as random search, grid search, ... as outlined in Section 4.2).

3. The model m is fitted on all $k - 1$ subsets other than s.

4. The model is applied on subset s and evaluated by computing a validation score.

5. The count of s is increased by 1 and steps 1-4 are repeated until $s = k$

After this procedure, k values of the parameter α and the associated validation scores can be observed and used to choose the optimal parameter. Through this validation procedure, models optimized with cross-validation usually do very well in predicting out-of-sample.

A.1.3 Model Development

For OLS, Table A3 and A4 depict the regional-specific coefficients of the second-step regressions for the first-round data, where Tables A5 and A6 contain the states/provinces of each region. As outlined in Section 4.2, we optimize the hyperparameters for the penalized regressions, the random forests and the neural networks. For each of the parameters of each method, we define a choice space within which the algorithm operates. That means, for a given method, it takes the choice space and tries out different hyperparameters. The hyperparameter is chosen which yields the lowest error rate on the validation set. The choice spaces for all our methods are depicted in the second column from the left of Tables A7 to A9. In the columns on the right, we have depicted the hyperparameters chosen by *hyperas*. Table A7 depicts choice space and hyperparameters for the penalized regressions, Table A8 for the random forests and Table A9 for the neural networks.

Table A3: Second-Step OLS Regressions - India

	India					
Variable	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
eduoldest	0.0115***	0.0197***	0.0103***	0.00699***		0.00528***
	(0.00197)	(0.00262)	(0.00156)	(0.00229)		(0.00191)
nfarm	0.0300***		0.0222***	0.0468***	0.0237***	0.0266***
	(0.00511)		(0.00524)	(0.00670)	(0.00601)	(0.00585)
person	-0.186***	-0.289***	-0.178***	-0.177***	-0.217***	-0.180***
	(0.00706)	(0.0189)	(0.00744)	(0.00716)	(0.00901)	(0.00608)
buffalo	0.0401***		0.0308***	0.0378***	0.0250***	
	(0.00519)		(0.00657)	(0.0122)	(0.00658)	
bike	0.0489***	0.0925***	0.0482***	0.0518***		
	(0.0138)	(0.0224)	(0.0137)	(0.0136)		
motorbike	0.136***	0.122***	0.255***	0.162***	0.160***	0.110***
	(0.0166)	(0.0332)	(0.0226)	(0.0220)	(0.0186)	(0.0173)
fan	0.0633***	-0.0476	0.158***	0.104***	0.181***	0.108***
	(0.0174)	(0.0297)	(0.0172)	(0.0194)	(0.0212)	(0.0159)
telephone	0.151***	0.314***	0.102***	0.212***	0.142***	0.240***
	(0.0167)	(0.0388)	(0.0287)	(0.0242)	(0.0205)	(0.0181)
cell	0.222***	0.274***	0.201***	0.193***	0.221***	0.209***
	(0.0192)	(0.0503)	(0.0306)	(0.0277)	(0.0243)	(0.0218)
fridge	0.121***	0.0762**	0.0883***	0.246***	0.112***	0.149***
	(0.0165)	(0.0357)	(0.0290)	(0.0263)	(0.0214)	(0.0212)
sizeloan	6.11e-07***	1.56e-06***	1.68e-06***	2.12e-07***	3.68e-07***	6.31e-07***
	(6.33e-08)	(3.17e-07)	(1.31e-07)	(5.84e-08)	(7.11e-08)	(5.29e-08)
size	0.0380***		0.00903**	0.0255***	0.0451***	0.0165***
	(0.00390)		(0.00427)	(0.00434)	(0.00599)	(0.00467)
tv	0.117***	0.131***	0.135***	0.174***	0.0833***	0.0972***
	(0.0161)	(0.0288)	(0.0174)	(0.0192)	(0.0176)	(0.0142)
own2	0.175***	0.110***	0.160***	0.152***	0.0958***	0.132***
	(0.0227)	(0.0407)	(0.0284)	(0.0222)	(0.0236)	(0.0178)
floor	0.0264*				0.0541***	0.0675***
	(0.0152)				(0.0194)	(0.0152)
wall	0.0341**	0.0948***		0.0310**	0.0385**	
	(0.0165)	(0.0279)		(0.0149)	(0.0182)	
wood	-0.0359**		-0.140***		-0.0778***	-0.135***
	(0.0156)		(0.0210)		(0.0202)	(0.0168)
personsq	0.00577***	0.0133***	0.00588***	0.00528***	0.00847***	0.00513***
	(0.000372)	(0.00151)	(0.000407)	(0.000434)	(0.000573)	(0.000356)
govhh	0.0995***	0.0791***	0.156***	0.123***	0.0993***	0.103***
	(0.0127)	(0.0237)	(0.0156)	(0.0152)	(0.0136)	(0.0122)
caste2	0.0482***	0.0767**				
	(0.0141)	(0.0328)				
caste6	0.0719***			-0.0413*		
	(0.0206)			(0.0219)		
age	0.00741***	0.0199***	0.0106***	0.00791***		0.00933***
	(0.00273)	(0.00486)	(0.00292)	(0.00284)		(0.00278)
eduhead	0.00590***			0.00617***	0.0108***	0.00619***
	(0.00194)			(0.00234)	(0.00169)	(0.00197)
children	-0.0359***		-0.0218**		-0.0312***	
	(0.00953)		(0.00964)		(0.0107)	
dratio	-0.0328***		-0.0450***	-0.0369***	-0.0514***	-0.0370***
	(0.00721)		(0.00721)	(0.00763)	(0.00701)	(0.00664)
nschool	0.0212***	0.0169*	0.0172***	0.0241***	0.00759	0.0157***
	(0.00557)	(0.00891)	(0.00598)	(0.00577)	(0.00684)	(0.00591)
agesq	-5.77e-05**	-0.000187***	-8.52e-05***	-6.64e-05**		-8.85e-05***
	(2.64e-05)	(4.88e-05)	(2.91e-05)	(2.84e-05)		(2.76e-05)
widow	-0.107***		-0.141***			
	(0.0375)		(0.0379)			
married	-0.0876***		-0.109***			
	(0.0334)		(0.0332)			

Note: Table continued on next page.

	India					
Variable	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
docc2	0.131***		0.104***	0.101***		0.0930***
	(0.0278)		(0.0354)	(0.0302)		(0.0316)
docc3	0.164***	0.174**	0.223***	0.113**		0.0993*
	(0.0488)	(0.0684)	(0.0724)	(0.0521)		(0.0519)
docc7	-0.107***		-0.106***	-0.0627***	-0.0857***	-0.0801***
	(0.0223)		(0.0171)	(0.0170)	(0.0183)	(0.0146)
docc8	-0.0436**					
	(0.0187)					
docc9	-0.0941***		-0.0624***		-0.120***	
	(0.0205)		(0.0233)		(0.0323)	
docc10	-0.0427*			-0.122***		
	(0.0259)			(0.0282)		
urb		0.0678**		0.0846***	0.0510**	0.0390***
		(0.0269)		(0.0180)	(0.0206)	(0.0147)
cow		0.0627***	0.0116*	0.0178***	0.0146**	
		(0.0132)	(0.00614)	(0.00641)	(0.00673)	
elec		0.157***			0.0721***	
		(0.0317)			(0.0227)	
toilet		0.117***	0.0559***		0.0431**	0.0632***
		(0.0392)	(0.0181)		(0.0168)	(0.0152)
caste8		0.469***			0.193***	0.0935***
		(0.0868)			(0.0542)	(0.0253)
docc4		0.0659*		0.0497		
		(0.0372)		(0.0304)		
incfarm			7.40e-07***	1.08e-06**	4.64e-07***	4.53e-07***
			(2.81e-07)	(4.78e-07)	(1.39e-07)	(8.12e-08)
roof			0.0651***			
			(0.0155)			
caste3			-0.0560***	-0.0758***		
			(0.0158)	(0.0180)		
caste4			-0.0715***	-0.0721***		
			(0.0183)	(0.0194)		
caste5			-0.282***	-0.239***	-0.0936***	
			(0.0251)	(0.0249)	(0.0239)	
caste7			0.165**			-0.256**
			(0.0792)			(0.104)
kitchen2				0.0684***		
				(0.0133)		
dwater2				0.0785***	-0.190***	
				(0.0176)	(0.0593)	
dwater3					-0.164***	
					(0.0610)	
dwater4					-0.154**	
					(0.0619)	
dwater5					-0.595***	0.164***
					(0.0940)	(0.0501)
sex					-0.0593**	-0.00633
					(0.0233)	(0.0176)
docc5					-0.0975*	
					(0.0585)	
medical						-0.0435***
						(0.0115)
disabled						0.0855***
						(0.0130)
Constant	7.111***	7.012***	6.771***	6.284***	7.013***	6.929***
	(0.0767)	(0.131)	(0.0819)	(0.0762)	(0.0983)	(0.0917)
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,621	1,521	5,923	5,083	4,133	7,843
R-Squared	0.565	0.659	0.608	0.663	0.619	0.509

Note: Standard errors in parentheses.*p<0.1; **p<0.05; ***p<0.01.

Table A4: Second-Step OLS Regressions - Indonesia

	Indonesia		
Variable	Region 1	Region 2	Region 3
floor	0.214***	0.215***	0.239***
	(0.0536)	(0.0271)	(0.0554)
urb		0.00624	-0.0210
		(0.0288)	(0.0435)
elec		0.221**	
		(0.0948)	
borrow		0.0919**	0.214***
		(0.0374)	(0.0726)
person	-0.206***	-0.261***	-0.291***
	(0.0220)	(0.0133)	(0.0248)
children		0.0594***	
		(0.0226)	
personsq	0.00524***	0.00898***	0.0108***
	(0.00127)	(0.000809)	(0.00162)
dratio		-0.0437***	
		(0.0104)	
age		0.0219***	
		(0.00485)	
agesq	-9.81e-06	-0.000232***	
	(1.59e-05)	(4.83e-05)	
farmdummy		0.0529*	
		(0.0302)	
posyanduknowledge		-0.158***	
		(0.0324)	
sizepercapita	0.00315***	8.53e-05*	0.00296***
	(0.000590)	(4.39e-05)	(0.000533)
toilettype2	-0.134	-0.0686*	
	(0.0824)	(0.0415)	
watertype2	-0.166***	-0.0987***	
	(0.0435)	(0.0306)	
watertype3		-0.162***	
		(0.0545)	
watertype5		0.0979**	0.383***
		(0.0382)	(0.0670)
cooktype2		0.251***	
		(0.0326)	
cooktype5	0.554***	0.372***	
	(0.139)	(0.0589)	
owntype2		-0.131***	-0.0495
		(0.0302)	(0.0588)
marstattype4	-0.592***	-0.215***	
	(0.151)	(0.0731)	
marstattype5		-0.236***	
		(0.0394)	
edulevhead4		0.152***	
		(0.0321)	
nschool	0.0927***	0.0944***	0.0880***
	(0.0223)	(0.0162)	(0.0228)
edulevhhm4		0.148***	
		(0.0307)	
tvdum	0.206***	0.151***	0.291***
	(0.0513)	(0.0306)	(0.0484)
fridgeown		0.107***	
		(0.0367)	
fridgeused	0.274***	0.306***	0.396***
	(0.0548)	(0.0409)	(0.0506)

Note: Table continued on next page.

	Indonesia		
Variable	Region 1	Region 2	Region 3
wall	0.147***		
	(0.0477)		
toilettype4	-0.302**		
	(0.145)		
cooktype4	-0.334***		
	(0.0505)		
edulevhhm2	-0.202***		
	(0.0699)		
dratiodum	-0.231**		
	(0.111)		
sex			0.163***
			(0.0536)
farmsize			3.94e-06*
			(2.08e-06)
toilettype3			-0.256***
			(0.0555)
owntype3			0.329***
			(0.0829)
marstattype3			-0.329
			(0.259)
Constant	13.49***	12.67***	13.95***
	(0.869)	(0.147)	(0.816)
State FE	Yes	Yes	Yes
Observations	1,775	5,176	1,772
R-squared	0.396	0.428	0.456

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01.

Table A5: Regions of India for Second-Step OLS regressions

Region	States Included
Region 1	Chandigarh, Delhi, Haryana, Himachal Pradesh, Jammu & Kashmir, Punjab, Rajasthan
Region 2	Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim
Region 3	Chhatishgarh, Madhya Pradesh, Uttaranchal, Uttar Pradesh
Region 4	Bihar, Jharkhand, Orissa, West Bengal
Region 5	Daman & Diu, Dadra & Nagar, Haveli, Goa, Gujarat, Maharashtra
Region 6	Andhra Pradesh, Karnataka, Kerala, Pondicherry, Tamil Nadu

Table A6: Regions of Indonesia for Second-Step OLS regressions

Region	Provinces Included
Region 1	Bangka-Belitung, Kepulauan Riau, Lampung, Riau, Sumatera Barat, Sumatera Selatan, Sumatera Utara
Region 2	Banten, Jakarta Raya, Jawa Barat, Jawa Tengah, Jawa Timur, Yogyakarta
Region 3	Bali, Kalimantan Selatan, Kalimantan Tengah, Kalimantan Timur, Nusa Tenggara Barat, Sulawesi Barat, Sulawesi Selatan, Sulawesi Utara

Table A7: Hyperparameter Optimization - Penalized Regression

		India					Indonesia				
Hyperparameter	Grid Space	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Elastic net parameter (λ)	0 to 1, in 0.1 intervals	0.2	0.4	0.5	0.9	0.3	0.9	0.8	1	0.4	0.1

Note: The λ Parameter is determined by the *cvlasso* command, differing for each regional unit.
(1) Baseline, (2) Second Round, (3) Robustness Check: Half Poverty Line, (4) Robustness Check: Short Vector, (5) Time Stability
Sources: Own models, *cvlasso* Ahrens et al. (2018)

Table A8: Hyperparameter Optimization - Random Forest

		India					Indonesia				
Hyperparameter	Choice Space	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Error criterion	mean squared error, mean absolute error	mae	mae	mae	mae	mae	mae	mae	mae	mae	mae
Number of regression trees estimated	8 to 96	64	16	48	24	16	32	24	48	64	36
Maximum depth of regression trees	16 to *None*, where *None* = no limit	*None*	128	*None*	*None*	*None*	*None*	32	252	*None*	*None*
Minimum number of samples required to split an internal node	2 to 64	24	16	2	32	16	16	32	24	24	16
Minimum number of samples required to be in a leaf node	2 to 64	4	16	4	16	16	8	16	4	8	10
Number of variables considered to determine split	20 to 88	64	48	64	40	48	32	64	40	15	40

Note: Not all choice spaces have been applied for all models. The choice space only represents the space considered overall.
(1) Baseline, (2) Second Round, (3) Robustness Check: Half Poverty Line, (4) Robustness Check: Short Vector, (5) Time Stability
Sources: Own models, *scikit-learn* (Pedregosa et al., 2011)

Table A9: Hyperparameter Optimization - Neural Network (1/2)

		India				
Hyperparameter	Choice Space	(1)	(2)	(3)	(4)	(5)
Number of hidden layers	1 to 15	3	3	3	3	3
Number of neurons in the hidden layers	25 to 396	291, 32, 97	194, 66, 97	291, 32, 97	70, 66, 97	245, 140, 40
Activation functions	sigmoid, relu, leaky relu, tanh	sigmoid, sigmoid, relu	sigmoid, sigmoid, relu	sigmoid, sigmoid, relu	sigmoid, sigmoid, relu	relu, relu, relu
Optimizing algorithm	Adam, Adadelta, Adagrad	Adam	Adam	Adam	Adam	Adam
Training epochs	5 to 50	20	15	20	15	25
Loss function	mean squared error, mean absolute error, quantile loss, mean abs. perc. error, mean sq. log. error	mae	mae	mae	mae	quantile(0.75)
Number of samples for gradient update (batch size)	8 to 128	32	32	64	16	32
Dropout between hidden layers	0 to 1	after all: 2.7%, 1.2%, 28.9%	after all: 19.1%, 1.1%, 10.7%	after all: 2.7%, 1.2%, 28.9%	after all: 19.1%, 1.1%, 10.7%	after all: 42.4%, 23.9%, 49.9%
Batch normalization between hidden layers	Yes or no	none	none	none	none	after first 2

Note: Not all choice spaces have been applied for all models. The choice space only represents the space considered overall. Dropout rates are rounded.
(1) Baseline, (2) Second Round, (3) Robustness Check: Half Poverty Line, (4) Robustness Check: Short Vector, (5) Time Stability
Sources: Own models, *Keras* (Chollet et al., 2015)

Table A10: Hyperparameter Optimization - Neural Network (2/2)

		Indonesia				
Hyperparameter	Choice Space	(1)	(2)	(3)	(4)	(5)
Number of hidden layers	1 to 15	3	3	3	3	3
Number of neurons in the hidden layers	25 to 396	264, 264, 176	40, 240, 285	264, 264, 264	80, 80, 80	220, 66, 88
Activation functions	sigmoid, relu, leaky relu, tanh	sigmoid, sigmoid, relu	relu, relu, relu	relu, relu, relu	sigmoid, relu, relu	relu, relu, relu
Optimizing algorithm	Adam, Adadelta, Adagrad	Adam	Adadelta	Adam	Adam	Adam
Training epochs	5 to 50	20	20	20	15	10
Loss function	mean squared error, mean absolute error, quantile loss, mean abs. perc. error, mean sq. log. error	mae	mae	mae	quantile(0.6)	quantile(0.75)
Number of samples for gradient update (batch size)	8 to 128	16	16	32	16	32
Dropout between hidden layers	0 to 1	after all: 8.2%, 50.7%, 85.6%	after all: 29.0%, 27.1%, 75.9%	after all: 61.1%, 73.7%, 65.2%	after first 2: 16.6%, 6.6%	after all: 22.8%, 54.2%, 33.3%
Batch normalization between hidden layers	Yes or no	after first 1	after first 2	after first 1	none	after first 1

Note: Not all choice spaces have been applied for all models. The choice space only represents the space considered overall. Dropout rates are rounded.
(1) Baseline, (2) Second Round, (3) Robustness Check: Half Poverty Line, (4) Robustness Check: Short Vector, (5) Time Stability
Sources: Own models, *Keras* (Chollet et al., 2015)

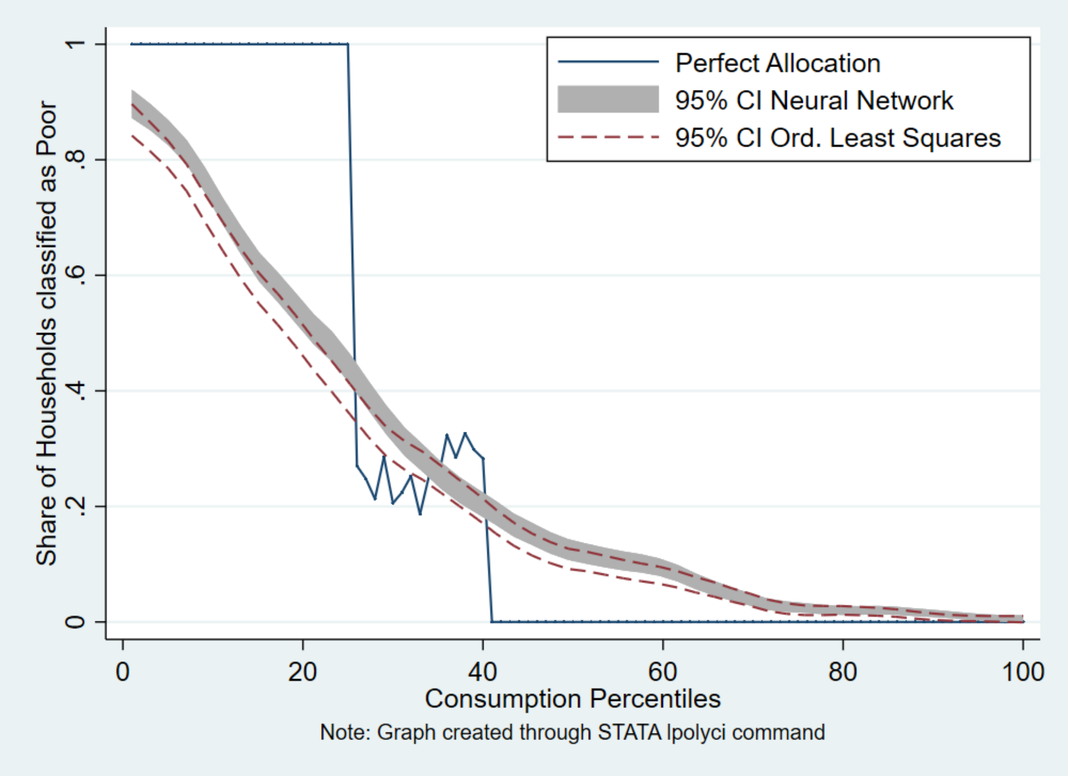
A.2 Results

In this subsection of the Appendix, we present supplementary graphs for our baseline results, all subgroup analyses but the baseline one, which can already be found in Section 5.1, and discuss our state and province analyses in detail. As all subgroup analyses are in line with the results from Section 5.1, we refrain from including them in the main body of the thesis. For consistency, we refer to Indonesian provinces from here on as states in order to use the same terminology as in the analyses for India.

A.2.1 Supplementary analysis - Baseline

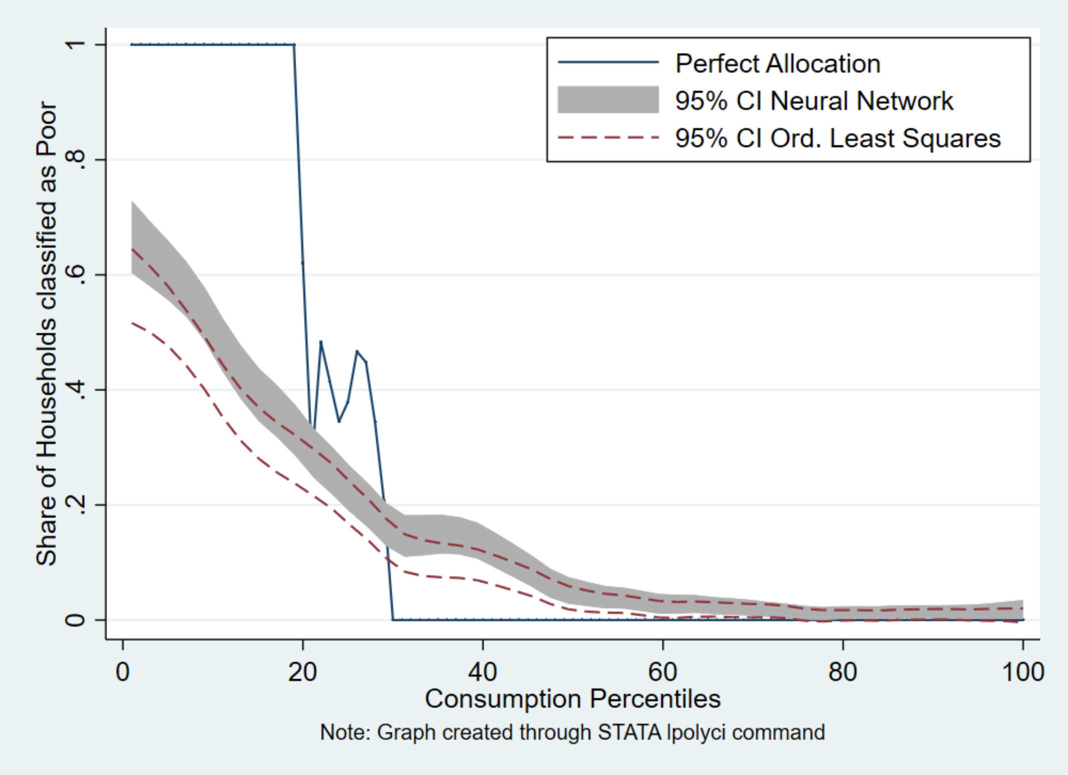
Figure A1 and A2 are supplementary graphs for our baseline analysis and correspond to Figures 4 and 5 in the main body of the thesis. In both graphs, we only depict the reference line, perfect classification and the targeting accuracy of the most precise methods, neural networks, and our benchmark, OLS. In order to analyze whether the smoothed graphs shown in Section 5.1 are actually significantly different from another, we plot only the 95%-confidence intervals of both methods. As can be seen in both figures, the confidence intervals of OLS, represented by the red, dashed lines, constantly overlap with the gray shaded areas that represent the confidence intervals of the neural networks. Hence, we conclude that although the neural network models seem to estimate consumption to be lower for all households throughout the consumption distribution compared to the other methods, the methods do not differ significantly.

Figure A1: Percentile Targeting OLS vs. NN (incl. 95% CI) - India, Baseline



Note: Graph created through STATA lpolyci command

Figure A2: Percentile Targeting OLS vs. NN (incl. 95% CI) - Indonesia, Baseline



Note: Graph created through STATA lpolyci command

A.2.2 Subgroup analysis - Second Round

Figure A3 and A4 depict the targeting accuracy for the different methods along the consumption distribution. All models follow the same patterns besides the random forests' which slightly overestimate consumption for both India and Indonesia. This leads to a line that lies below the other three as households are less likely to be categorized poor which also translates into the aggregate error rates in Table 5 where the random forest models have higher exclusion and lower inclusion error rates.

The subgroup analysis for urban vs. rural households and those with a female vs. male head is shown in Table A11. The only statistically significant coefficient on a 5%-level is the interaction between the urban dummy and the neural network. However, with a magnitude of 0.8 percentage points compared to a constant of 13.8 percent, the effect is economically not significant. Additionally, visual inspection of the corresponding maps, Figures A5 and A6, does not suggest that the methods target certain states differently accurate. For an econometric analysis on the different states, please see Section A.2.6. Overall, the subgroup analysis for the data sets of the second round is in line with our findings in Section 5.1 for the baseline.

Figure A3: Targeting along Consumption Percentiles - India, Second Round

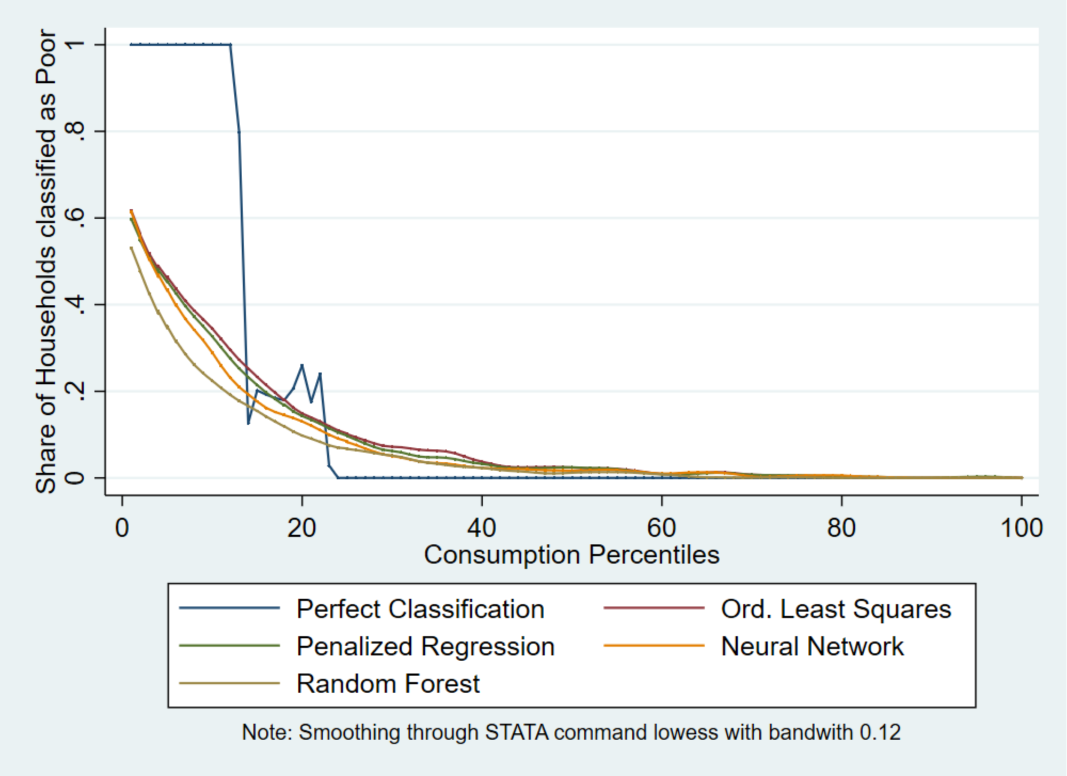


Figure A4: Targeting along Consumption Percentiles - Indonesia, Second Round

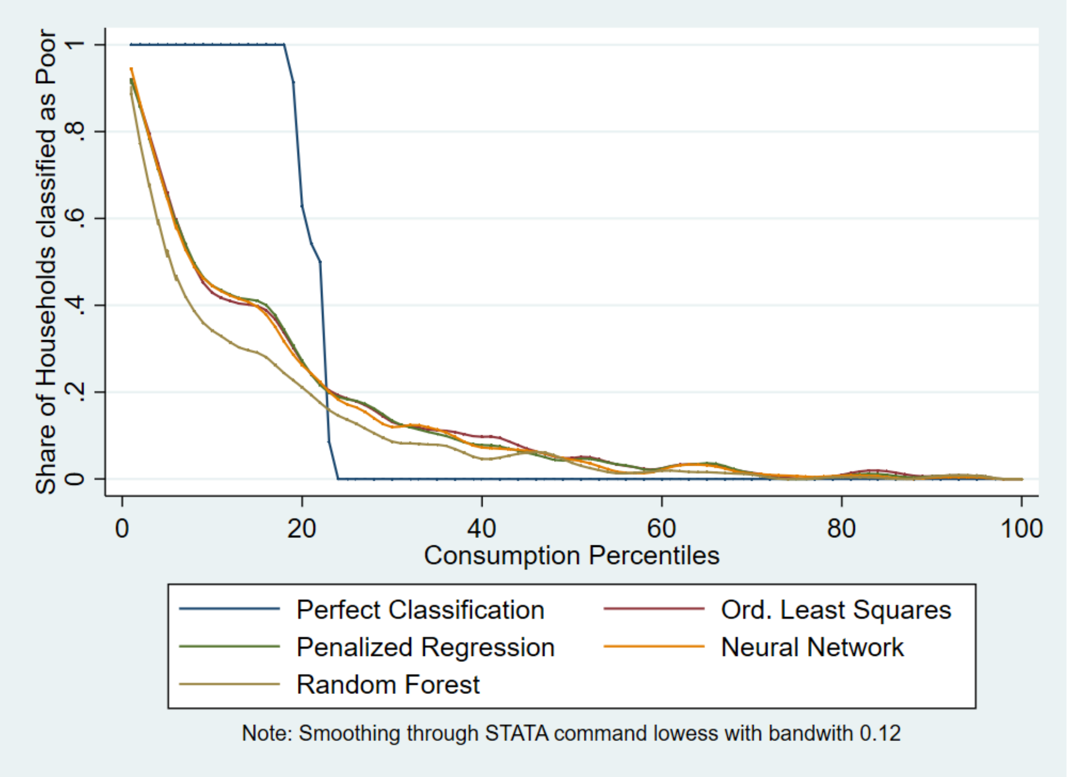
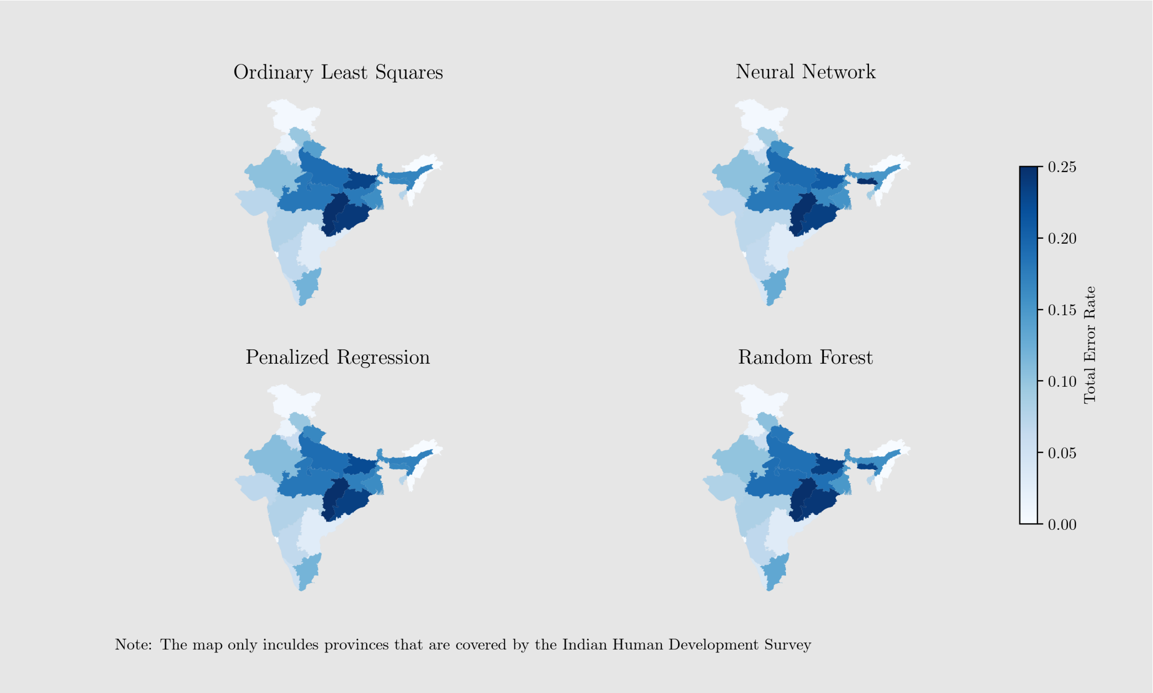


Table A11: Total Error Rates for Inspected Subgroups - Second Round

	India		Indonesia	
	(1)	(2)	(3)	(4)
Urban x PR	-0.002		0.008	
	(0.002)		(0.005)	
Urban x NN	-0.008^{***}		0.007	
	(0.003)		(0.010)	
Urban x RF	-0.004		0.006	
	(0.005)		(0.013)	
Female x PR		0.003		0.007
		(0.003)		(0.007)
Female x NN		0.007^{*}		-0.007
		(0.004)		(0.015)
Female x RF		-0.001		0.010
		(0.005)		(0.017)
Constant	0.138^{***}	0.119^{***}	0.174^{***}	0.136^{***}
	(0.007)	(0.005)	(0.012)	(0.008)
Method FE	Yes	Yes	Yes	Yes
Urban FE	Yes	No	Yes	No
Female FE	No	Yes	No	Yes
Observations	41,480	41,480	14,056	14,056

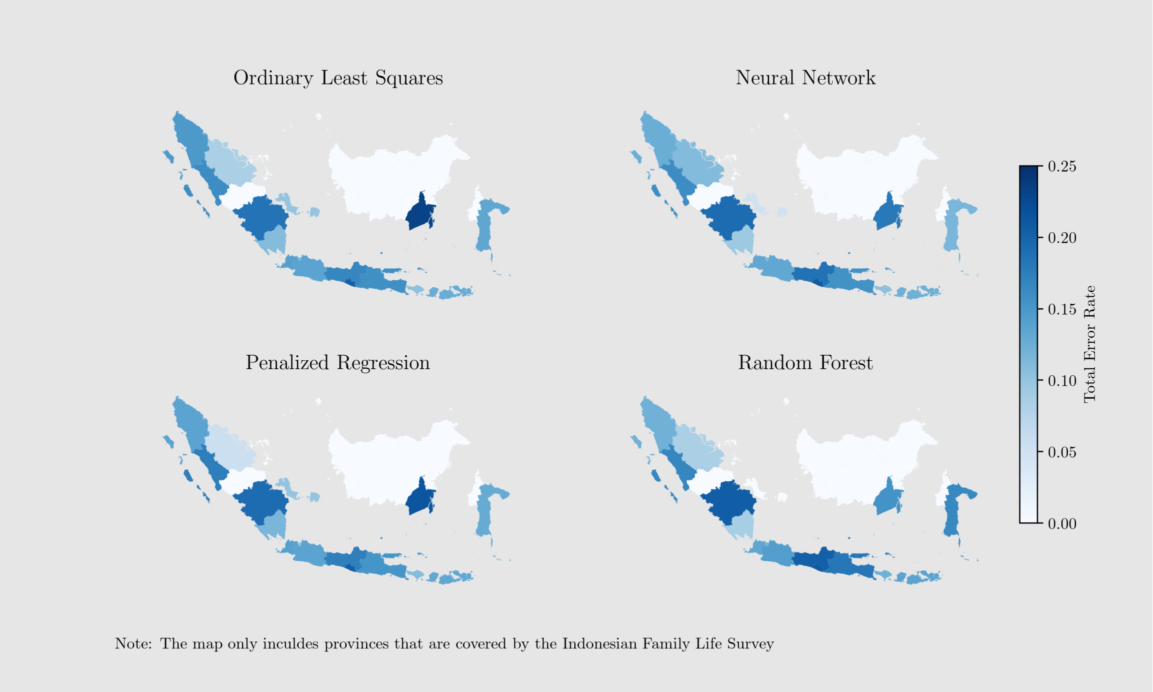
Note: Standard errors are clustered at the district level.
*p<0.1; **p<0.05; ***p<0.01
(1) and (3): Urban vs. Rural Households
(2) and (4): Female vs. Male Head of Household
PR: Penalized Regression, NN: Neural Network, RF: Random Forest

Figure A5: Total Error Rates across States - India, Second Round



Note: The map only inculdes provinces that are covered by the Indian Human Development Survey

Figure A6: Total Error Rates across Provinces - Indonesia, Second Round



Note: The map only inculdes provinces that are covered by the Indonesian Family Life Survey

A.2.3 Subgroup analysis - Rob. Check: Half Poverty Line

Figure A7 and A8 depict the targeting accuracy for the different methods along the consumption distribution. Due to the very low poverty lines that only classify around 5 percent of the households each country as poor, all models follow the perfect classification line for most of the consumption distribution. The two neural network models seem to be the most precise in terms of classifying the poor correctly, however, even they often classify more than half of the poorest percentiles as non-poor. However, as shown in Table 6 in the main body, the differences are not significant on aggregate for the total error rate.

The subgroup analysis for urban vs. rural households and those with a female vs. male head is shown in Table A12 and does not include any interaction that is significant on a 5%-level. Additionally, visual inspection of the corresponding maps, Figures A9 and A10, does not indicate that the methods target certain states differently accurate. It is interesting to note however, that the error rates for Indian states, despite being small on aggregate, can be substantial for some states such as Chhattisgarh. For an econometric analysis on the different states, please see Section A.2.6. Overall, the subgroup analysis for our robustness check using a smaller poverty line confirms our findings for the baseline in Section 5.1.

Figure A7: Targeting along Consumption Percentiles - India, Half Poverty Line

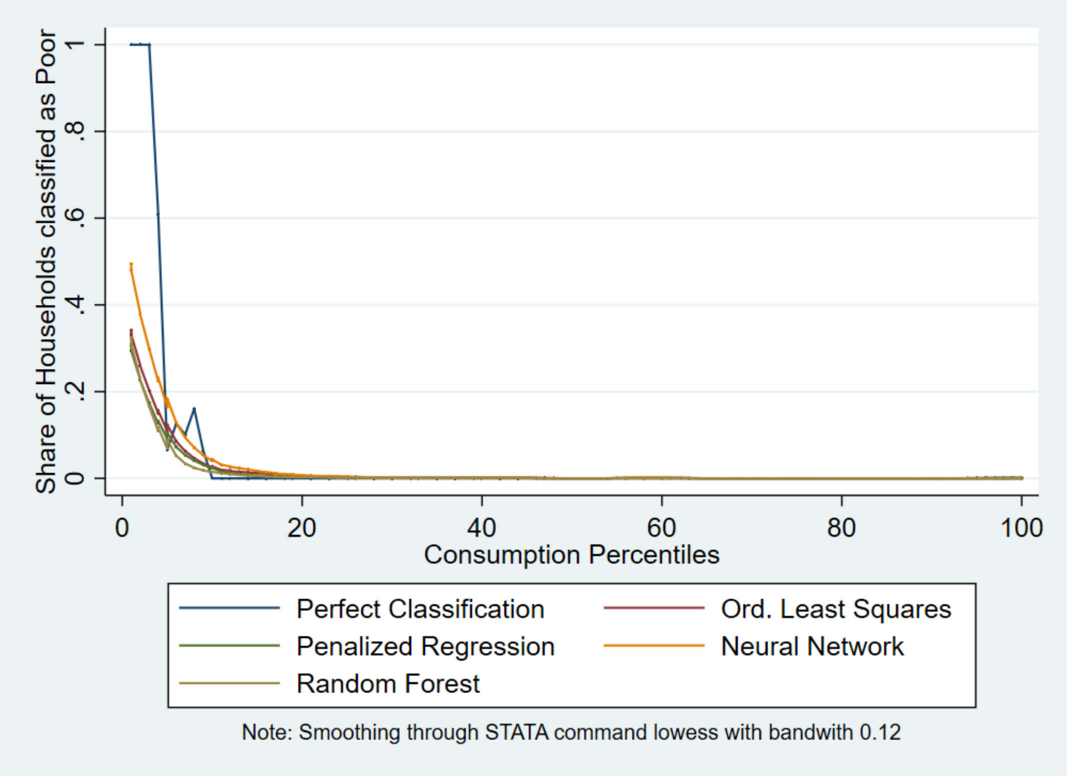


Figure A8: Targeting along Consumption Percentiles - Indonesia, Half Poverty Line

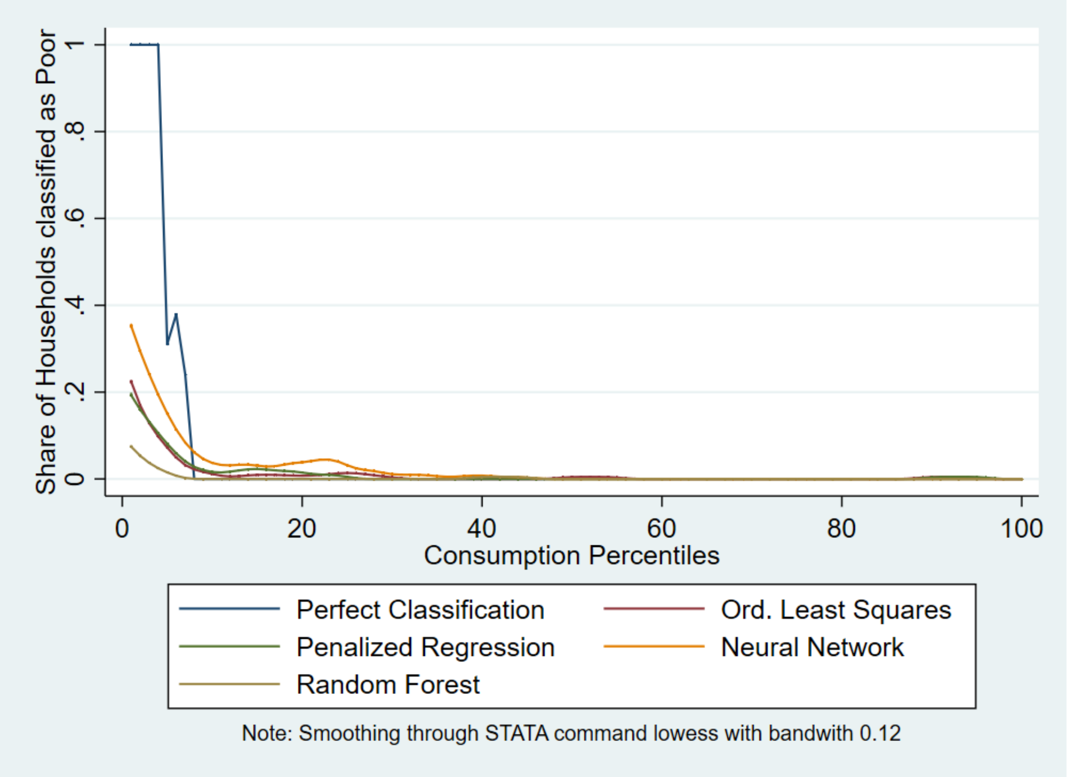
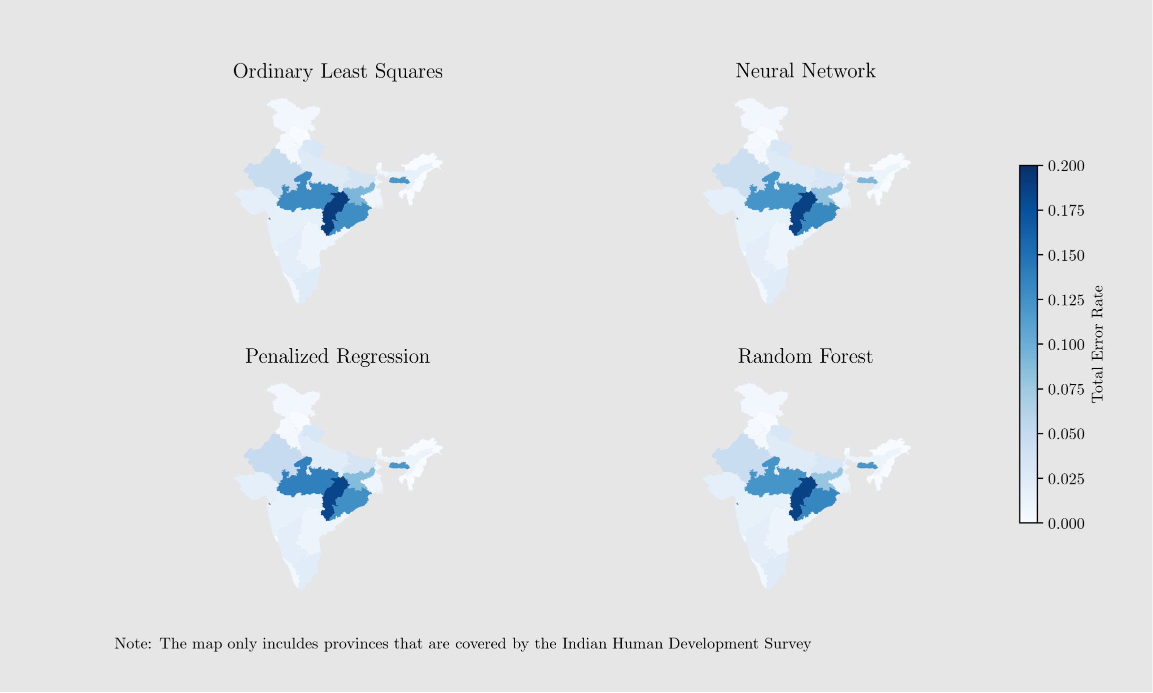


Table A12: Total Error Rates for Inspected Subgroups - Half Poverty Line

	India		Indonesia	
	(1)	(2)	(3)	(4)
Urban x PR	-0.001^{*}		0.000	
	(0.001)		(0.002)	
Urban x NN	0.002		0.005	
	(0.002)		(0.004)	
Urban x RF	0.002		-0.004	
	(0.002)		(0.004)	
Female x PR		0.001		-0.005
		(0.001)		(0.005)
Female x NN		0.0004		-0.001
		(0.002)		(0.009)
Female x RF		-0.00002		-0.003
		(0.002)		(0.005)
Constant	0.047^{***}	0.038^{***}	0.061^{***}	0.0367^{***}
	(0.004)	(0.003)	(0.005)	(0.004)
Method FE	Yes	Yes	Yes	Yes
Urban FE	Yes	No	Yes	No
Female FE	No	Yes	No	Yes
Observations	41,484	41,484	11,632	11,632

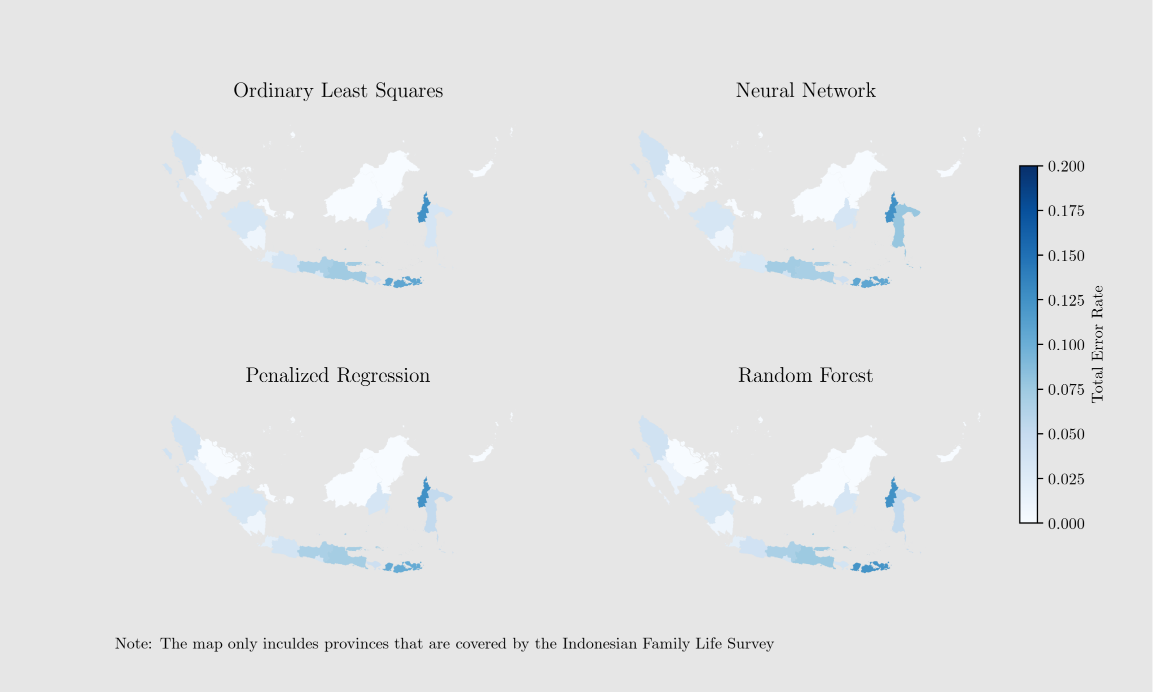
Note: Standard errors are clustered at the district level.
*p<0.1; **p<0.05; ***p<0.01
(1) and (3): Urban vs. Rural Households
(2) and (4): Female vs. Male Head of Household
PR: Penalized Regression, NN: Neural Network, RF: Random Forest

Figure A9: Total Error Rates across States - India, Half Poverty Line



Note: The map only inculdes provinces that are covered by the Indian Human Development Survey

Figure A10: Total Error Rates across Provinces - Indonesia, Half Poverty Line



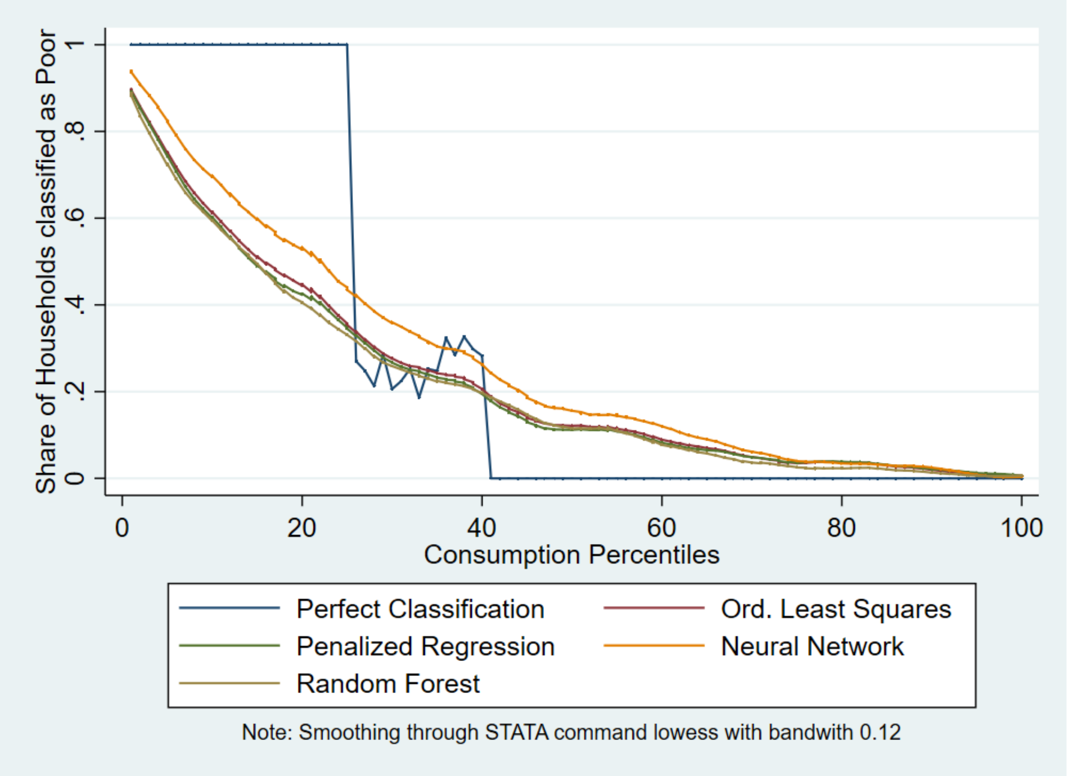
Note: The map only inculdes provinces that are covered by the Indonesian Family Life Survey

A.2.4 Subgroup analysis - Rob. Check: Short Vector

Figure A11 and A12 depict the targeting accuracy for the different methods along the consumption distribution. All models follow the same patterns besides the neural networks' which slightly underestimate consumption for both India and Indonesia. This leads to a line that lies above the other three as households are more likely to be categorized poor which translates into the aggregate error rates in Table 7 where the neural network models have lower exclusion and higher inclusion error rates.

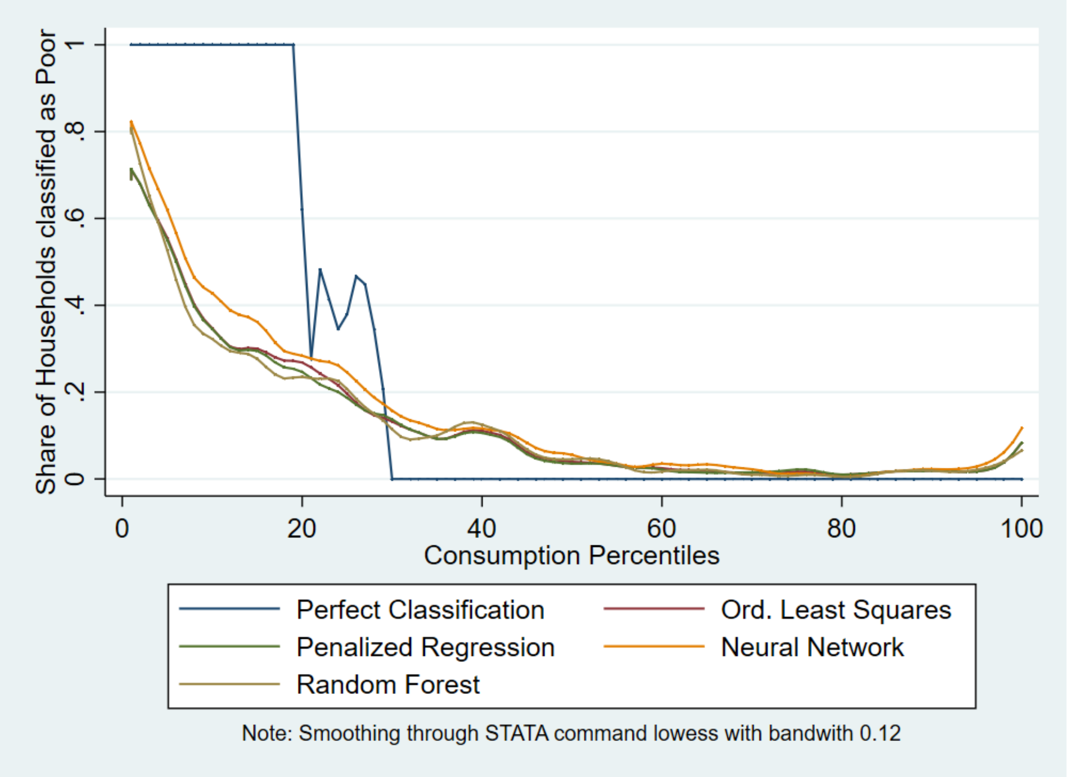
The subgroup analysis for urban vs. rural households and those with a female vs. male head is shown in Table A13 and does not include any interaction that is significant on a 5%-level. Additionally, visual inspection of the corresponding maps, Figures A13 and A14, does not indicate that a single method targets certain states differently accurate compared to the others. One small difference that can be noted is that in the case of India, the models of OLS and penalized regression, target the states of Madhya Pradesh and Bihar slightly worse. We investigate whether this could be a systematic difference in an econometric analysis together with the other analyses in Section A.2.6. Overall, the subgroup analysis for our robustness check using a short set of input variables confirms our findings for the baseline in Section 5.1.

Figure A11: Targeting along Consumption Percentiles - India, Short Vector



Note: Smoothing through STATA command lowess with bandwith 0.12

Figure A12: Targeting along Consumption Percentiles - Indonesia, Short Vector



Note: Smoothing through STATA command lowess with bandwith 0.12

Table A13: Total Error Rates for Inspected Subgroups - Short Vector

	India		Indonesia	
	(1)	(2)	(3)	(4)
Urban x PR	−0.003		0.003	
	(0.003)		(0.003)	
Urban x NN	0.008		0.004	
	(0.005)		(0.008)	
Urban x RF	0.003		−0.002	
	(0.006)		(0.011)	
Female x PR		0.005		0.004
		(0.004)		(0.003)
Female x NN		0.002		−0.011
		(0.008)		(0.010)
Female x RF		0.006		−0.009
		(0.010)		(0.008)
Constant	0.211^{***}	0.193^{***}	0.207^{***}	0.165^{***}
	(0.007)	(0.008)	(0.009)	(0.007)
Method FE	Yes	Yes	Yes	Yes
Urban FE	Yes	No	Yes	No
Female FE	No	Yes	No	Yes
Observations	41,484	41,484	11,632	11,632

Note: Standard errors are clustered at the district level.
*p<0.1; **p<0.05; ***p<0.01
(1) and (3): Urban vs. Rural Households
(2) and (4): Female vs. Male Head of Household
PR: Penalized Regression, NN: Neural Network, RF: Random Forest

Figure A13: Total Error Rates across States - India, Short Vector

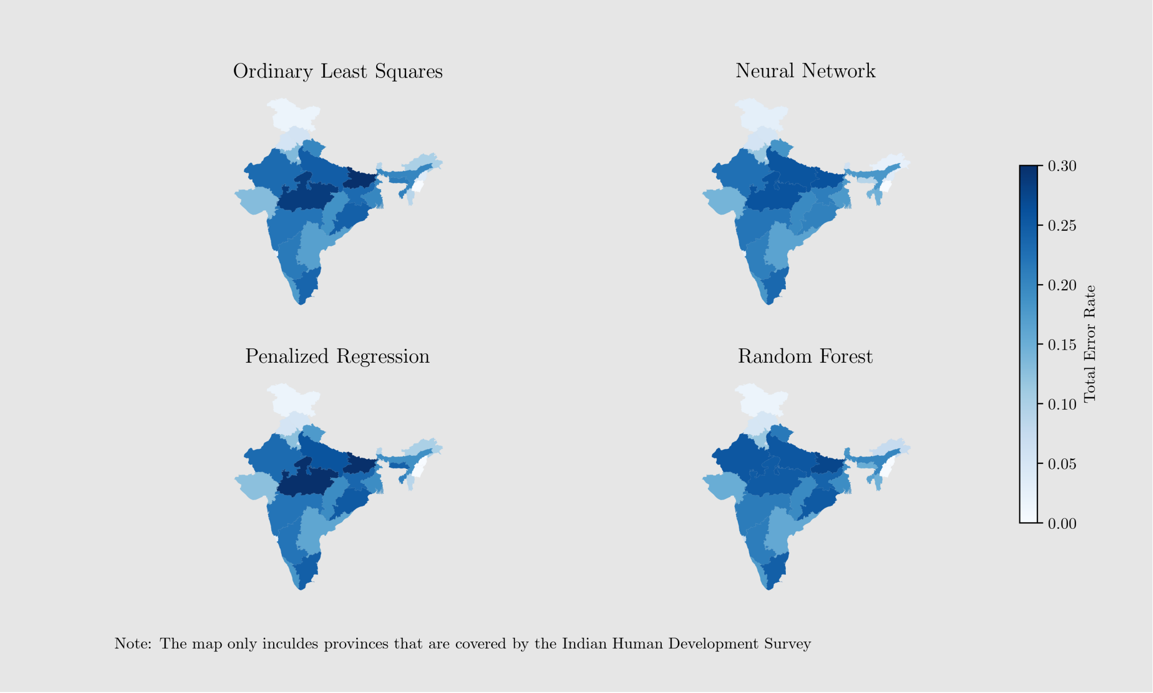
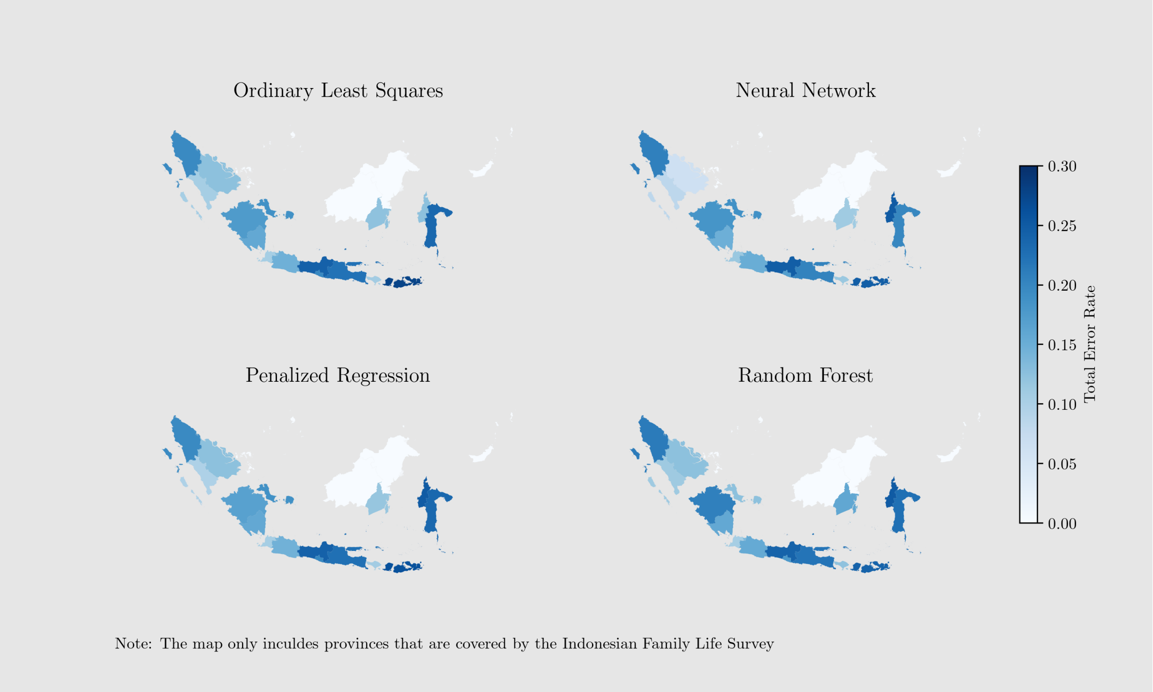


Figure A14: Total Error Rates across Provinces - Indonesia, Short Vector

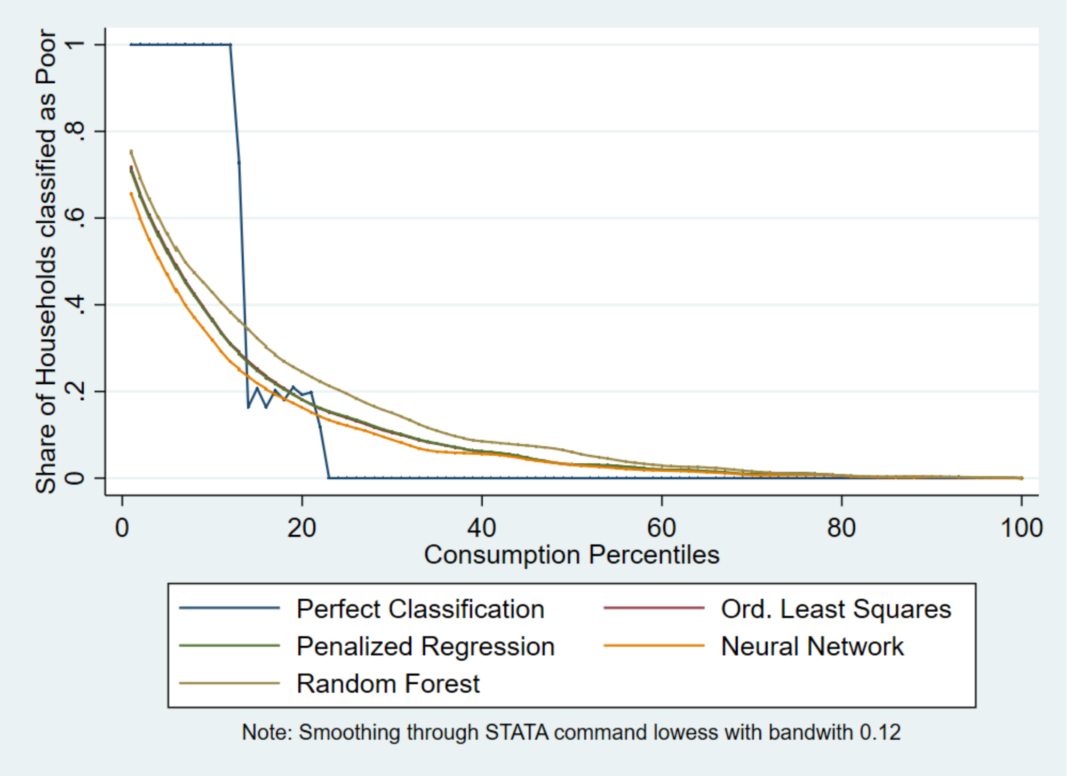


A.2.5 Subgroup analysis - Time Stability

Figure A15 and A16 depict the targeting accuracy for the different methods along the consumption distribution. All models follow the same patterns besides the neural networks' which slightly overestimate consumption for Indonesia. This leads to a line that lies below the other three as households are less likely to be categorized poor. This translates into the aggregate error rates in Table 9 where the neural network models have higher exclusion and lower inclusion error rates, evening out in the total error rate that is not statistically significantly different.

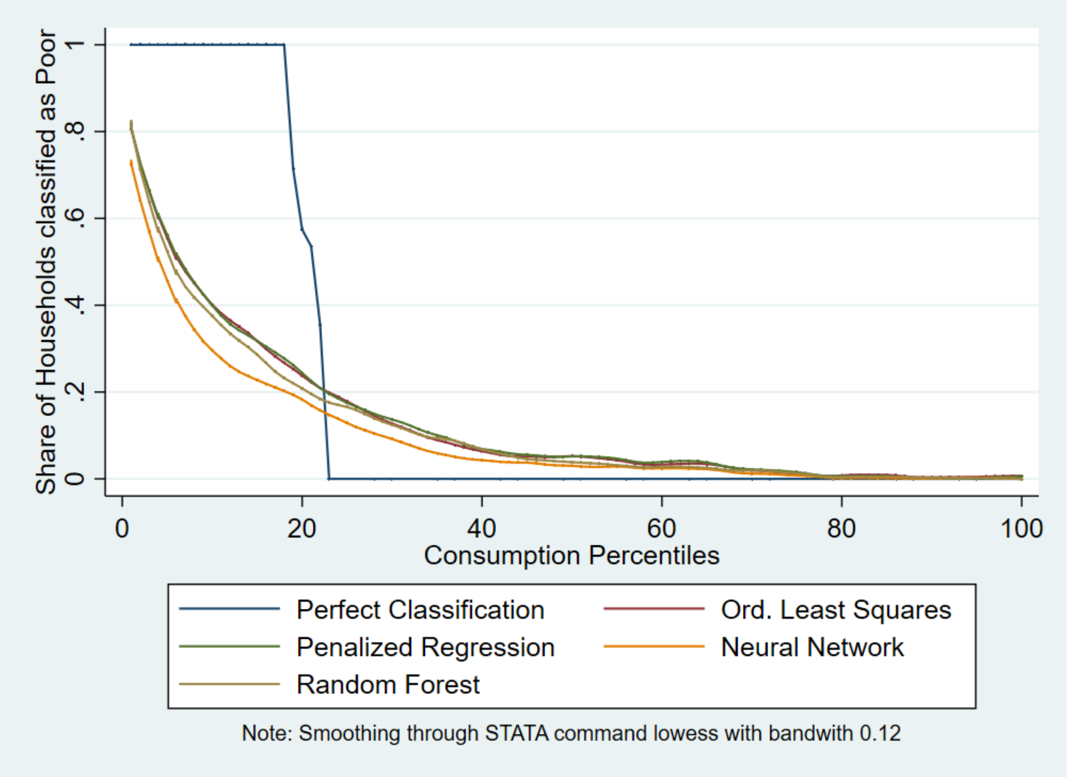
The subgroup analysis for urban vs. rural households and those with a female vs. male head is shown in Table A14. The only statistically significant coefficient on a 5%-level is the interaction of the urban dummy and the neural network in the Indonesia data set. However, with a magnitude of 1.4 percentage points compared to a constant of 19.1 percent, the effect is arguably economically not significant. Additionally, visual inspection of the corresponding maps, Figures A17 and A18, does not indicate that the a single method targets certain states differently accurate compared to the others. Note however, that both machine learning tools seem to target the small Indonesian state Sulawesi Barat badly. We investigate whether this could be a systematic difference in an econometric analysis together with the other analyses in Section A.2.6. Overall, the subgroup analysis for our time stability setting confirms our findings for the baseline in Section 5.1.

Figure A15: Targeting along Consumption Percentiles - India, Time Stability



Note: Smoothing through STATA command lowess with bandwith 0.12

Figure A16: Targeting along Consumption Percentiles - Indonesia, Time Stability



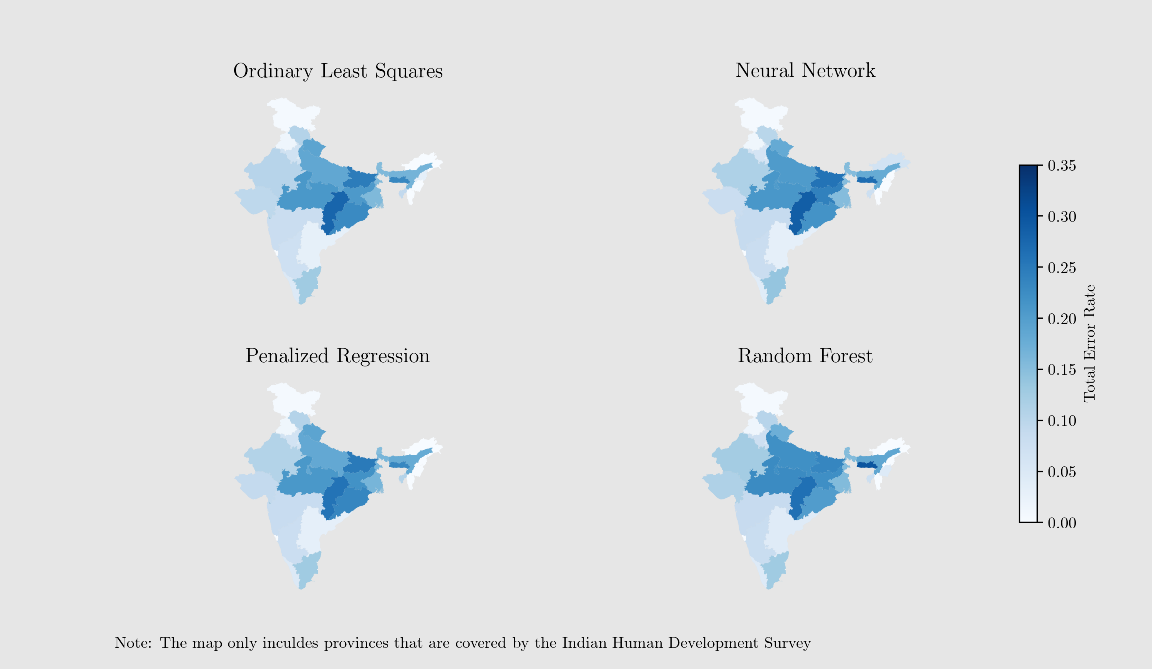
Note: Smoothing through STATA command lowess with bandwith 0.12

Table A14: Total Error Rates for Inspected Subgroups - Time Stability

	India		Indonesia	
	(1)	(2)	(3)	(4)
Urban x PR	-0.001		-0.004	
	(0.001)		(0.003)	
Urban x NN	0.0003		-0.014^{***}	
	(0.002)		(0.005)	
Urban x RF	0.004		0.002	
	(0.003)		(0.006)	
Female x PR		-0.001		-0.002
		(0.001)		(0.005)
Female x NN		0.001		-0.001
		(0.003)		(0.008)
Female x RF		-0.007^{*}		-0.002
		(0.004)		(0.007)
Constant	0.145^{***}	0.123^{***}	0.191^{***}	0.145^{***}
	(0.005)	(0.005)	(0.007)	(0.005)
Method FE	Yes	Yes	Yes	Yes
Urban FE	Yes	No	Yes	No
Female FE	No	Yes	No	Yes
Observations	165,964	165,964	56,224	56,224

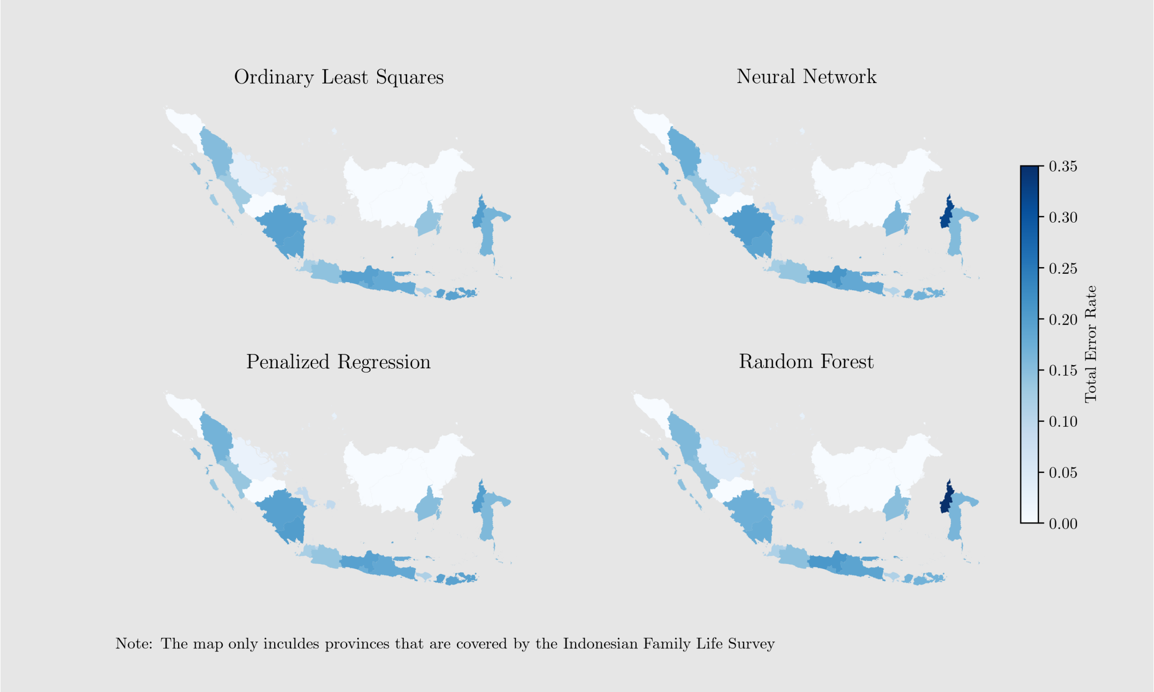
Note: Standard errors are clustered at the district level.
*p<0.1; **p<0.05; ***p<0.01
(1) and (3): Urban vs. Rural Households
(2) and (4): Female vs. Male Head of Household
PR: Penalized Regression, NN: Neural Network, RF: Random Forest

Figure A17: Total Error Rates across States - India, Time Stability



Note: The map only inculdes provinces that are covered by the Indian Human Development Survey

Figure A18: Total Error Rates across Provinces - Indonesia, Time Stability



Note: The map only inculdes provinces that are covered by the Indonesian Family Life Survey

A.2.6 State and Province analysis

The maps, which depict the total error rates in the different states and provinces, have given little proof that our four methods target certain states differently compared to each other. While there is variance in the total error rate across states, no state seems to be consistently targeted worse or better by one method across all analyses. To confirm this visual evidence, we analyze the total error rates in a regression setting. We bundle all but the six most populous states in each country together to obtain large enough sample sizes and then check using the regression depicted in equation 4 for differences in means across the largest states and the different methods. The results for all five of our analyses are depicted in Table A15 and A16.

For India, we see that there are between zero and three coefficients per analyses that are significant at a 5%-level. For example, for our robustness check using a poverty line at 50 percent of the official level, the penalized regression model leads to 1 percentage point higher total error rates in West Bengal compared to the other methods and all other states that are not specifically included as dummies in the regression. Although this might seem small in size, it represents an increase of more than 25 percent from the mean value of 3.7 percent. Hence, it could be considered economically significant. However, while we observe some similar significant coefficients across the different analyses, they are not robust across all analyses.

This pattern is mirrored in the results for the state analyses on the Indonesian data sets. There are between zero and four statistically significant coefficients in the analysis, at a 5%-level, with the robustness check using a short vector being the one with the most ones. However, there is no interaction term for which the coefficient is constantly significant across all analyses. Hence, we conclude that the econometric analysis supports our hypothesis of no systematic discrimination against certain states by any particular method.

Table A15: Test for Differences across the Largest States - India

	India				
	(1)	(2)	(3)	(4)	(5)
PR x Uttar Pradesh	0.009	0.001	−0.002	0.010^{*}	0.002
	(0.008)	(0.003)	(0.002)	(0.006)	(0.003)
PR x Maharashtra	0.016	−0.007	0.001	0.009^{*}	−0.00001
	(0.010)	(0.008)	(0.001)	(0.005)	(0.004)
PR x Bihar	0.002	0.003	0.001	−0.006	0.002
	(0.007)	(0.004)	(0.001)	(0.005)	(0.002)
PR x West Bengal	0.015^{*}	0.001	0.010^{***}	0.020^{***}	0.003
	(0.009)	(0.003)	(0.003)	(0.007)	(0.003)
PR x Madhya Pradesh	0.002	0.001	−0.001	0.0005	0.0003
	(0.006)	(0.004)	(0.002)	(0.004)	(0.002)
PR x Tamil Nadu	0.018^{**}	−0.001	−0.001	0.008	0.002
	(0.007)	(0.008)	(0.002)	(0.008)	(0.002)
NN x Uttar Pradesh	0.007	0.001	0.001	0.016	0.012^{**}
	(0.010)	(0.008)	(0.002)	(0.014)	(0.005)
NN x Maharashtra	-0.027^{**}	−0.019	−0.002	-0.033^{*}	0.012
	(0.013)	(0.022)	(0.003)	(0.017)	(0.011)
NN x Bihar	−0.021	0.001	0.003	−0.015	0.007
	(0.015)	(0.011)	(0.002)	(0.015)	(0.008)
NN x West Bengal	-0.021^{*}	0.005	−0.006	−0.018	0.005
	(0.011)	(0.008)	(0.010)	(0.016)	(0.006)
NN x Madhya Pradesh	0.011	−0.004	−0.0001	0.006	0.006^{*}
	(0.013)	(0.009)	(0.004)	(0.011)	(0.003)
NN x Tamil Nadu	0.008	0.008	−0.001	0.003	0.009^{***}
	(0.010)	(0.010)	(0.002)	(0.012)	(0.003)
RF x Uttar Pradesh	0.009	−0.005	−0.001	0.003	0.035^{***}
	(0.011)	(0.008)	(0.002)	(0.012)	(0.009)
RF x Maharashtra	−0.008	0.001	−0.002	-0.029^{*}	-0.018^{*}
	(0.017)	(0.020)	(0.003)	(0.017)	(0.011)
RF x Bihar	−0.015	−0.011	0.002	−0.010	−0.003
	(0.013)	(0.014)	(0.002)	(0.018)	(0.005)
RF x West Bengal	−0.013	0.007	−0.007	-0.038^{***}	0.005
	(0.017)	(0.011)	(0.011)	(0.014)	(0.006)
RF x Madhya Pradesh	0.003	0.002	−0.001	−0.012	−0.004
	(0.011)	(0.007)	(0.002)	(0.011)	(0.004)
RF x Tamil Nadu	0.027^{*}	0.011	0.0005	0.004	0.002
	(0.015)	(0.010)	(0.001)	(0.010)	(0.003)
Constant	0.148^{***}	0.096^{***}	0.037^{***}	0.163^{***}	0.097^{***}
	(0.008)	(0.005)	(0.004)	(0.008)	(0.004)
Method FE	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes
Observations	41,484	41,480	41,484	41,484	165,964

Note: Standard errors are clustered at the district level. *p<0.1; **p<0.05; ***p<0.01
(1) Baseline, (2) Second Round, (3) Robustness Check: Half Poverty Line, (4) Robustness Check: Short Vector, (5) Time Stability.
Six chosen states counting for 37.5 percent of the population for (1),(3),(4).
Six chosen states counting for 38.8 percent of the population for (2).
Six chosen states counting for 38.6 of population for (5).
PR: Penalized Regression, NN: Neural Network, RF: Random Forest

Table A16: Test for Differences across the Largest Provinces - Indonesia

	Indonesia				
	(1)	(2)	(3)	(4)	(5)
PR x Jawa Barat	0.003	−0.013	−0.001	0.004*	0.013**
	(0.012)	(0.017)	(0.003)	(0.002)	(0.006)
PR x Jawa Timur	0.018	−0.015*	−0.001	0.009*	−0.003
	(0.013)	(0.008)	(0.003)	(0.006)	(0.004)
PR x Jawa Tengah	−0.003	−0.001	−0.001	0.002	−0.007
	(0.010)	(0.007)	(0.004)	(0.003)	(0.005)
PR x Sumatera Utara	−0.017	0.006	−0.001	0.004	−0.003
	(0.014)	(0.008)	(0.003)	(0.005)	(0.006)
PR x Banten	−0.004	−0.005	−0.003	0.007**	0.001
	(0.016)	(0.008)	(0.007)	(0.003)	(0.005)
PR x Jakarta Raya	0.003	−0.001	−0.001	0.004*	−0.004
	(0.008)	(0.005)	(0.003)	(0.002)	(0.005)
NN x Jawa Barat	0.001	−0.015	−0.005	0.027***	0.025***
	(0.020)	(0.018)	(0.004)	(0.008)	(0.007)
NN x Jawa Timur	0.002	0.004	−0.005	0.030**	0.003
	(0.019)	(0.012)	(0.004)	(0.014)	(0.007)
NN x Jawa Tengah	0.007	0.004	−0.011**	0.019	−0.005
	(0.016)	(0.013)	(0.006)	(0.014)	(0.008)
NN x Sumatera Utara	0.024	0.027**	0.001	0.018	0.018*
	(0.021)	(0.013)	(0.009)	(0.016)	(0.010)
NN x Banten	−0.024	0.006	−0.010	−0.003	0.004
	(0.018)	(0.010)	(0.007)	(0.009)	(0.008)
NN x Jakarta Raya	0.007	−0.0002	−0.005	0.026**	0.002
	(0.028)	(0.020)	(0.004)	(0.011)	(0.011)
RF x Jawa Barat	−0.009	−0.026	−0.003	0.016	0.010
	(0.017)	(0.019)	(0.003)	(0.021)	(0.009)
RF x Jawa Timur	−0.009	0.005	−0.003	−0.001	0.010
	(0.012)	(0.016)	(0.003)	(0.018)	(0.007)
RF x Jawa Tengah	−0.005	0.006	−0.001	0.007	0.006
	(0.014)	(0.011)	(0.005)	(0.010)	(0.008)
RF x Sumatera Utara	0.015	0.037**	−0.003	−0.004	0.017***
	(0.023)	(0.016)	(0.003)	(0.020)	(0.006)
RF x Banten	−0.011	0.026*	−0.001	−0.003	0.014*
	(0.015)	(0.015)	(0.007)	(0.017)	(0.008)
RF x Jakarta Raya	−0.019	−0.008	−0.003	−0.001	−0.004
	(0.019)	(0.027)	(0.003)	(0.027)	(0.013)
Constant	0.156***	0.149***	0.042***	0.174***	0.157***
	(0.011)	(0.009)	(0.006)	(0.007)	(0.007)
Method FE	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes
Observations	11,632	14,056	11,632	11,632	56,224

Note: Standard errors are clustered at the district level. *p<0.1; **p<0.05; ***p<0.01
(1) Baseline, (2) Second Round, (3) Robustness Check: Half Poverty Line, (4) Robustness Check: Short Vector, (5) Time Stability.
Six chosen states counting for 59.8 percent of the population for (1),(3),(4).
Six chosen states counting for 57.6 percent of the population for (2).
Six chosen states counting for 58.1 of population for (5).
PR: Penalized Regression, NN: Neural Network, RF: Random Forest