Thailand and The World Economy
P-ISSN: 2630-0931
E-ISSN: 2651-0529
Homepage : https://www.tci-thaijo.org/index.php/TER

Vol. 38, No. 1, January - April 2020 Page [91-116]

Better Model Selection for Poverty Targeting through Machine Learning: A Case Study in Thailand

Pisacha Kambuya[*]
Division of Macroeconomic and Data Analysis,
Fiscal Policy Research Institute Foundation, Thailand

Received 11 April 2019, Received in revised form 24 January 2020,
Accepted 25 January 2020, Available online 16 March 2020

Abstract

The proxy means test (PMT) is a method for targeting poor households that should obtain benefits from social programs. The PMT estimates income or expenditure by the ordinary least square (OLS) method using a set of variables that is correlated with welfare measurements because income and expenditure are difficult to measure directly. Variable selection in the OLS process requires stepwise regression which is a time-consuming task when the set of variables is very large. This study proposed the Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF) algorithms which are a part of the Machine Learning field to improve variable selection and performance of the PMT model by focusing on the out-of-sample targeting accuracy of poor households in Thailand. Data for this study were taken from the 2016 Thailand Socio-Economic Survey (SES). Results showed that PMTs based on the selected variables from RF reduced the number of poor households that are classified as non-poor households (an exclusion error) and increased the poverty accuracy rate (target poor households as poor accurately) at national, urban and rural levels. However, the inclusion error was still high. Performances of PMTs based on the selected variable from stepwise regression and LASSO were similar. PMTs with stepwise regression and LASSO selected variables outperformed RF selected variables in terms of inclusion error reduction. On the other hand, an exclusion error for PMTs based on RF selected variables was shown to be significantly less than PMTs using stepwise regression and LASSO selected variables. Since there is a trade-off between inclusion and exclusion errors, results of this study suggested that if the objective of the social welfare program is to help the poor, then PMTs based on the variable selection of RF would be more appropriate.

Keywords: Proxy Means Test, Poverty Targeting, Machine Learning, Variable Selection, LASSO, Random Forest
JEL Classifications: C14, C52, I32, I38

[*] Address: 1175/2 Krungthep-Nonthaburi 39, Krungthep-Nonthaburi Rd., Wong Sawang, Bang Sue, Bangkok 10800, Thailand. Email: pisacha@fispri.org

1. Introduction

Over the past few years, the Government of Thailand has launched many social programs or universal schemes to provide monetary subsidies to the poor. One such scheme is the subsistence allowance program for older people while another well-known social program is the Register for State Welfare. This was set up under the responsibility of the Ministry of Finance in 2016 and provides a subsidy of 300 baht per month to people who are not employed or have an income of less than 30,000 baht per year to purchase goods through a welfare card known as “Bat Sawasdikarn Hang Rat”. This scheme is classified as a means-testing program to target the poor based on income criteria and asset ownership. The advantage of mean-testing scheme is to identify the actual poor who should obtain priority in the program. These schemes are compared to universal schemes that cannot reduce inequality and increase the fiscal burden. To achieve high targeting accuracy through a means-testing approach, minimizing inclusion errors (the non-poor are identified as poor) and exclusion errors (the truly poor are classified as non-poor) is paramount. In terms of an impact on poverty, the program implementers should focus on reducing the exclusion error, while budget constraint aspects concern the alleviation of inclusion error. Hence, tools for targeting the poor are needed to consider these two types of error.

One popular method to target the poor is called the proxy means test (PMT). This method is based on the assumption that measurements of household consumption expenditure and income are inappropriate. They are difficult to obtain directly as some households or individuals underreport their income or expenditure. Therefore, here, the estimation of household income or consumption was implemented by a linear regression model using household characteristics as a proxy such as age, quality of the dwelling, ownership of farmland and durable goods, or educational level of the household head as the explanatory variables. Variables that significantly correlated with an income or expenditure were considered as the selected variables in the model. The PMT was used to create effective outcomes for poverty targeting among all targeting methods in Latin America (M. Grosh & Baker, 1995). Nowadays, PMT has become the common tool for targeting the poor in several social programs because full means tests are costly and time-consuming to monitor.

The PMT is a tool that can be utilized to quickly and easily target poor households. However, the OLS method requires time to conduct both variable selection and the process of running and comparing the performance of several models over a large set of variables. Stepwise regression is required to perform these tasks. Random Forest (RF) and Least Absolute Shrinkage and Selection Operator (LASSO) are algorithms in the field of machine learning. These powerful predictive models can perform variable selection to enhance prediction accuracy and statistical interpretability. Studies by McBride and Nichols (2015, 2016) and Sohnesen and Stender (2017) have shown that both RF and LASSO can reduce an exclusion error in PMTs to more accurately target poor households. Here, RF and LASSO were applied to select the variables for building a PMT model that can reduce an exclusion error.

This article is organized as follows. Section 2 assesses the current literature on poverty targeting by considering an alternative algorithm for model and variable selection. Section 3 describes the frameworks of the LASSO and RF algorithms, while Section 4 sheds light on the data and empirical methodology. Section 5 presents and discusses the results with conclusions drawn in Section 6.

2. Review of the Literature

2.1 Development of Proxy Means Test (PMT)

Ravallion (1996) designed and constructed poverty targeting in terms of regression of individual poverty which he measured by using the variety of a household's characteristics as a proxy to predict an income or expenditure. One advantage of poverty regression is that policymakers gain knowledge concerning which region, A or B, should get priority in any social program. A PMT based on an ordinary least square (OLS) regression model using the log of income or expenditure as the dependent variable has become a common tool to target the poor in developing countries. For example, Ahmed and Bouis (2002) implemented an OLS regression model for constructing a PMT to target the needy for a food subsidy program in Egypt using per capita consumption expenditure as a welfare measurement. This study used the household size, education of members and ownership of durable goods as a proxy. Results showed that the government could reduce the budgetary allocation by about 74 percent which was more than the saving from the selected practical model. However, the OLS regression method for PMT presents two problems. First, the OLS method minimizes the sum of the squares between the true and predicted outcomes and is different from the minimized poverty problem. Second, variables on the right-hand side of the equation (explanatory variables) face an endogenous problem (M. Grosh & Baker, 1995).

Most studies of PMT classified variables that were correlated with an income or expenditure into several categories such as the household demographics, ownership of assets, characteristics of dwelling, education of household head and location variables. Thus, the question is raised as to which method they used to select these variables as the best indicator in the final model. For a large set of candidate variables, stepwise regression is preferred for selection in a PMT tool (Brown, Ravallion, & Van De Walle, 2016; Nguyen & Lo, 2016; Narayan & Yoshida, 2005; M. Grosh & Baker, 1995). However, James and McCulloch (1990) suggested that the stepwise regression procedure cannot rank or provide the best variables based on their importance.

Over the past decade, several studies have proposed alternative methods besides OLS regression to improve the robustness of PMT models. Quantile regression was suggested by Koenker and Bassett Jr (1978). This method has more robust outliers than the OLS model, while Houssou et al. (2007) argued that the OLS method was more robust than the quantile regression method. In addition to OLS regression, they also employed the linear probability model (LPM), probit analysis and quantile regression methods to test the robustness and out-of-sample validity of the model. Results showed that quantile regression performed with moderate accuracy for in-sample predictions of poverty but was less robust, while the OLS method and probit analysis performed better for out-of-sample predictions, suggesting that the probit method provided optimized accuracy and robustness for the PMT model.

For implementing PMT in Thailand, the Child Support Grant Program (2015) used the PMT to target eligibility of newborns and pregnant women in poor households for receiving the grant under the following five conditions as having household monthly income lower than 3,000 baht per person, having a dependency member, housing conditions, not owning a car or truck and farmers owning less than one rai of land (about 1,600 square metres). Another program that used the PMT for targeting the poor was the grant for poor students in Thailand. Punyasavatsut (2017) employed the same conditions as the Child Support Grant Program; however, poor students at the provincial level were targeted while the Child Support Grant Program was implemented at the national level.

2.2 Variable Selection

An important step of the PMT is variable selection. Questions concerning which variables should be included in the final model for PMT are an interesting topic. For PMT with OLS regression, stepwise regression is used to select a set of variables by eliminating variables that are not statistically significant with a dependent variable and also do not decrease the explanatory power of the model (R-squared) when they are added or excluded. In practice, the set of variables to be used in a PMT should be easy for staff to collect and then calculate the PMT score. Therefore, small sets of variables are better than too many variables. However, stepwise regression is a time-consuming task for OLS when the set of variables is large and also has an endogeneity problem. Many studies have proposed alternative algorithms in the machine learning field to study variable selection and try to capture data patterns to understand the variables in the dataset. This is especially true for non-linear variables that linear regression techniques such as OLS cannot capture directly.

Tibshirani (1996) first introduced the Least Absolute Shrinkage and Selection Operator (LASSO) method as a machine learning algorithm that proposes the penalized term called “loss function” to regularize (shrink) the coefficient in the OLS estimator to zero for a variable that provides less correlation with a dependent variable. Thus, LASSO renders the OLS as a spare model. In other words, LASSO can eliminate variables by shrinking those variables to zero inside its algorithm to obtain a small set of variables. The coefficient from LASSO will converge or diverge from the coefficient of OLS dependent on the lambda parameter that is chosen to ensure that LASSO coefficients have a high bias from OLS. Belloni and Chernozhukov (2013) suggested the OLS-post LASSO method by proposing that LASSO selects variables and model at the first step and then estimates these LASSO selected variables using OLS. They described their results by deriving the theoretical properties of post-model selection of the LASSO estimator. Results showed that if LASSO can capture the true variables in the model, then OLS will make the error smaller than that proposed only by LASSO. The performance of OLS-post LASSO in terms of a convergence coefficient close to zero is as good as LASSO. Similarly, a study by Hastie et al. (2015) also provided an example of LASSO as the first step for variable selection and then proposed OLS. The result showed that proposing OLS-post LASSO induced additional sparsity in the model.

However, LASSO fails in terms of missing the true variable that provides the main effect on a dependent variable. Random Forest or RF (Breiman, 2001) provides a variable selection within its algorithm that can be used to estimate variable importance. RF runs an algorithm based on the aggregate bootstrap, growing several trees and then predicting the result by averaging the outcomes from each tree. This method can be used to reduce the variance and improve the accuracy of the model. The RF algorithm also performs better in out-of-sample prediction and can capture non-linear variables. RF ranks the order of selected variables using the importance of each variable. High importance value means that the variable has a high effect on a dependent variable. If the high importance variables are excluded, the accuracy of the model will decrease. However, the variable importance of RF cannot provide important information such as the coefficients of linear models. This leads to the crucial problem of the inability to obtain and realize the magnitude of the main effect of the selected variable using RF. Hence, Random Forest is known as a non-interpreted model.

2.3 Improvement of PMT through Machine Learning

In the poverty prediction literature, the application of RF is still scant and very recent. Otok and Seftiana (2014) determined that the RF method was very accurate in identifying poor households who were eligible for social assistance packages in Indonesia, while Thoplan (2014) used an RF method to predict poverty in Mauritius. Results showed that the RF model provided the most accurate prediction for poverty. Using the PMT, McBride and Nichols (2015, 2016) were successful in applying the RF method to predict targeting performance compared with the linear regression-based models to improve the targeting accuracy of PMT. They used the World Bank's Living Standards Measurement Study (LSMS) survey data. Their study compared the out-of-sample targeting accuracy in Malawi, Bolivia and Timor-Leste. Results revealed that quantile RF was better at estimating poor households as poor or undercoverage rate declined while the leakage was still high. They concluded that the RF method could significantly improve out-of-sample performance by between 2 and 18 percent.

In their recent study, Sohnesen and Stender (2017) used the LASSO and RF methods to predict poverty using one year of data for prediction within the same year and two years of data to predict poverty over time. Their results indicated that RF was a good predictor for poverty and provided a more robust estimate than linear regression methods. The RF model provided a highly accurate poverty prediction in both urban and rural areas but did not offer a more accurate prediction compared with LASSO and linear regression models at the national level. However, RF was proven to predict poverty with accuracy, even though a small number of selected variables were used in the model instead of a full set of variables. This study concluded that the RF method was simple and easy to use. Furthermore, Kshirsagar et al. (2017) used a bootstrap LASSO to select a subset of variables that provided an accurate prediction of poverty rate, while Knippenberg et al. (2017) captured the food insecurity dynamics of households using the Coping Strategy Index (CSI) as a measurement to implement LASSO and RF algorithms to choose the ten best-selected variables. Their results indicated that the predictive accuracy of CSI between LASSO and RF methods was similar since LASSO provided greater accuracy than RF by only 0.8 percent.

The literature reviews revealed that previous social programs in Thailand were mostly individually provided. Punyasavatsut (2017) proposed the PMT to target poor students by using the household dataset since a student is a dependent person and income or expenditure should be evaluated from a parent or household members, while studies by Otok and Seftiana (2014) and McBride and Nichols (2015, 2016) proposed the PMT to target poor households. This study conducted the PMT to target the poor as a household unit.

3. Algorithm Frameworks

3.1 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO was developed by Tibshirani (1996) based on the least square estimator by the addition of a penalty term. The LASSO estimator can be shown as equation (3.1).

$$\min_{\beta_0,\beta_j}\left\{\frac{1}{N}\sum\nolimits_{i=1}^{N}(y_i-\beta_0-\sum\nolimits_{j=1}^{P}x_{ij}\beta_j)^2\right\}, \tag{3.1}$$

$$\text{subject to } \sum\nolimits_{j=1}^{P}\left|\beta_j\right|\le t\ , \tag{3.2}$$

Equation (3.1) is an optimization problem in terms of least square functional form with a subjective equation (3.2), where $i=1,...,N$ denotes the number of observations, $j=1,...,P$ denotes the number of explanatory variables and t is the parameter that defines a regularization size.

To obtain β_{LASSO}, the function of equation (3.1) aims to optimize the problem by minimizing the residual sum of squares (RSS). The LASSO estimator, β_{LASSO}, can be solved by equation (3.3)

$$\beta_{LASSO}=\min_{\beta_0,\beta_j}\left\{\frac{1}{N}\sum\nolimits_{i=1}^{N}(y_i-\beta_0-\sum\nolimits_{j=1}^{P}x_{ij}\beta_j)^2+\lambda\sum\nolimits_{j=1}^{P}\left|\beta_j\right|\right\}, \tag{3.3}$$

where the second term is "l_1 loss function". This loss function is summed across all absolute coefficients and then multiplied by lambda (λ), which is the parameter that defines the Bayesian shrinkage degree of the problem. β_{LASSO} is selected to minimize RSS and the LASSO estimator then allows us to tune the lambda. In other words, the residual sum of square (RSS) increases or decreases depending on the size of lambda. This is the advantage of LASSO to reduce an error by tuning the lambda parameter.

There is no theory that supports the choice of lambda. For the relationship between lambda and coefficient, if the lambda converges to zero ($\lambda\to 0$) the objective function then becomes an OLS estimator and β_{LASSO} is equal to β_{OLS}. However, if the lambda value is positive, then the coefficient of β_{LASSO} will divert from the coefficient of β_{OLS}. Moreover, if the lambda converges towards infinity ($\lambda\to\infty$) the coefficients of β_{LASSO} will tend to close to zero. In other words, the coefficients will have been shrunk to zero. Therefore, all coefficient estimates depend on the chosen value of lambda.

In practice, the lambda is chosen through a cross-validation (CV) method. Initially, using the untransformed coefficients to ensure that the value of lambda will be between the mean of zero ($\lambda.\min$) and standard deviation of one ($\lambda.1se$).

The LASSO estimator can select the variable by penalizing the model based on the sum of an absolute value of coefficients. Some variables will be zero after optimizing the objective function and the coefficients that remain non-zero will be considered as the variable selection.

3.2 Random Forest

Random Forest (RF) was first introduced by Ho (1995). She proposed stochastic modeling to construct decision tree-based classifiers which could be randomly expanded to increase accuracy for training and testing (unseen) data. In other words, this method constructs multiple trees in a random feature subspace (set of variables). Amit and Geman (1997) then studied this new approach that aimed to shape classification and illustrate performance in high dimensions in terms of the number of shaped classes and the degree of variability within classes. They defined a large number of geometric arrangements in the split at each node, based on the growing binary classification of trees.

The Random Forest algorithm grows the trees based on the decision tree that is used to predict the outcome in terms of the Classification and Regression Trees (CART) procedure. This is one class of supervised learning methods as a machine learning algorithm that predicts observations from the data in terms of characteristics (classification trees) and continued variables (regression trees) which split a space into regions following the binary decision rule. This study sheds light on the regression trees model, in particular, the Random Forests for making predictions of household expenditure.

Regression tree models are constructed by building a tree. Each node follows the recursive binary tree as a splitting algorithm as follows (Hastie et al., 2009):

The algorithm decides on the splitting variable, X_j. The splitting point, $X_j = s$, then defines the half planes of R_1 and R_2 which can be shown as:

$$R_1(j,s) = \{X \mid X_j \le s\} \text{ and } R_2(j,s) = \{X \mid X_j > s\}, \tag{3.4}$$

then select X_j and s to solve the minimization problem,

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right], \tag{3.5}$$

Any X_j and s can be solved by equation (3.6), where c_i is the average of y_i in each R_i region,

$$\hat{c}_1 = n^{-1}\textstyle\sum_i (y_i \mid x_i \in R_1(j,s)) \text{ and } \hat{c}_2 = n^{-1}\textstyle\sum_i (y_i \mid x_i \in R_2(j,s)). \tag{3.6}$$

For the best split, this algorithm divides the data into the two results of the region and repeats the splitting process at each of the two regions. Then, repeat on all of the resulting regions. The optimal size of the growing tree depends on the data. Very large trees will be confronted with an over-fitting problem, while small-sized trees cannot capture the structure (under-fitting problem). The algorithm stops the process when each branch meets the terminal node.

However, one problem of the regression trees model is that a small change in data can affect the split of the trees and high variance. The error can spread from the top of a tree down. To alleviate the variance, bootstrap aggregation or bagging (Breiman, 1996) is used.

To increase the prediction accuracy of a model with low-variance, bagging builds the prediction models separately $\hat{f}(x), \hat{f}^{2}(x), \ldots, \hat{f}^{B}(x)$ on B separate training datasets and then averages the resulting predictions. This generates a new training set using random bootstrap sampling to replace an original dataset. The set of tree models can then be trained independently by applying the regression tree algorithm on the new training dataset. The predicted responses are calculated by averaging all the models $\hat{f}^{*b}(x)$, which can be written as:

$$\hat{f}_{bag}(x) = B^{-1}\sum_{b=1}^{B}\hat{f}^{*b}(x)\,. \tag{3.7}$$

Unfortunately, even though the variance is reduced, the constant term of variance remains. The idea is that a set of B identical distribution and regression trees are correlated with variance, σ^{2} . As an example, let ρ represent the pairwise correlation between the trees. Then the average set of B independent observations is $\rho\sigma^{2} + \frac{(1-\rho)}{B}\sigma^{2}$. The $\frac{(1-\rho)}{B}\sigma^{2}$ term will converge towards zero as B grows large but the term $\rho\sigma^{2}$ still persists (McBride & Nichols, 2016). An extension of RF was proposed by (Breiman, 2001). This version of RF reduces variance by using bagging to improve the classification accuracy by combining the resulting classifications of randomly generated training sets. The out-of-bag (OOB) method was also implemented to gain accuracy in the model by measuring the generalization error (or out-of-sample error). In other words, measuring how the accuracy of an algorithm can predict outcome values for unseen data. Avoiding the over-fitting problem can minimize the generalization error.

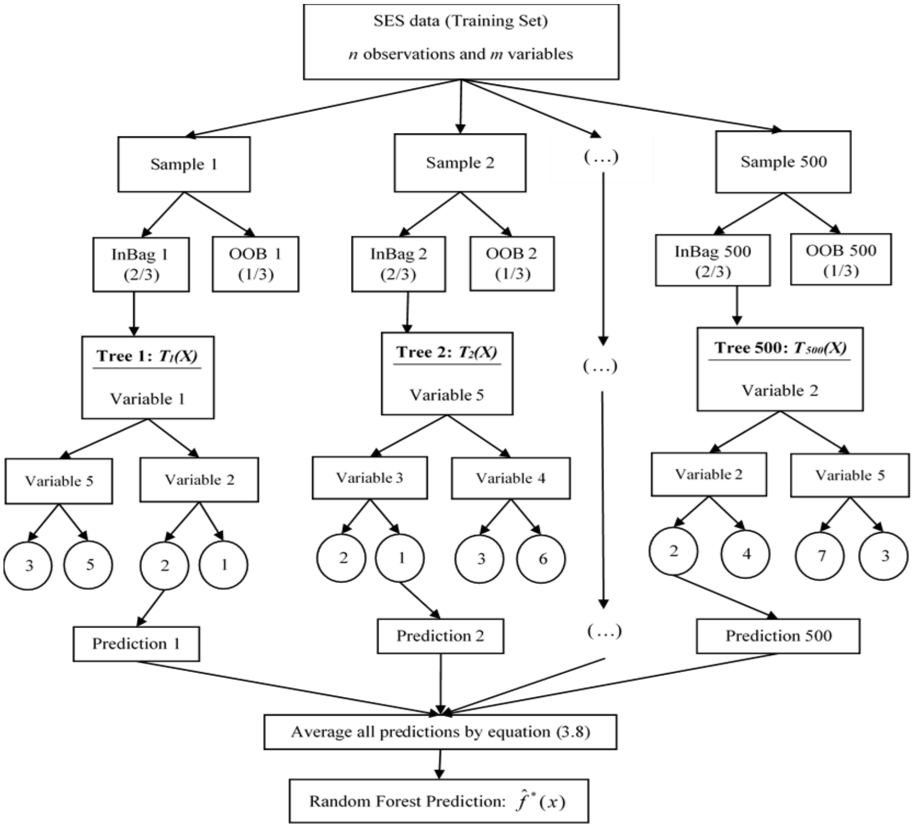
Random Forest RF is closely related to the bagging method by containing a large number of decision trees on bootstrapped training samples. Every time RF splits the tree, the process begins with the prediction of a single tree, $B = \{T_{1}(X), \ldots, T_{B}(X)\}$, where $X = \{x_{1}, \ldots, x_{M}\}$ is the full set of M-dimensional vectors of predictors (independent variables). Then, randomly sampling the m predictors from this full set.

Ensemble produces b outputs, $\{\hat{f}_{1}(x) = T_{1}(X), \ldots, \hat{f}_{B}(x) = T_{B}(X)\}$, where $\hat{f}_{b}(x)$, b = 1, …, B is the prediction of training data by the b^{th} tree. Outputs of all trees are aggregated to perform one final prediction, $\hat{f}_{b}^{*}(x)$. Thus, $\hat{f}^{*}(x)$ is the class predicted by the majority of trees in the classification problem and the average of individual tree predictions for a regression problem. Then, the Random Forest predictor is constructed in equation (3.8) as:

$$\hat{f}^{*}(x) = B^{-1}\sum_{b=1}^{B}\{T_{1}(X), \ldots, T_{B}(X)\}\,. \tag{3.8}$$

Figure 1 illustrates how the RF learns in a context of this study. Assuming we have a SES data of size n, RF will feed each tree a sample size n with replacement. The RF builds a number 500 of regression trees making them grow from different training data (In-bag) subsets by randomly resampling 2/3 of full dataset with replacement. Hence, most data will be used multiple times in different models. The model which was trained in an in-bag data will be measured error in the OOB. On the other hand, when the RF makes a tree grow, for example, it uses the best predictor within a subset of variables ($m = 3$) which has been selected randomly from the overall set of input variables ($M = 5$). These special characteristics of RF confer a greater prediction stability and accuracy, at the same time, avoid the correlation of the different regression tress, increase the diversity of patterns that can be learnt from data. The multiple predictions of all 500 regression trees for a given vector used as training are then averaged to obtain a unique estimation of the monthly per capita consumption expenditure.

Figure 1: Random Forest Process



Source: Author’s summarization.

4. Data and Empirical Methodology

4.1 Data

The monthly per capita consumption expenditure data in this study comes from the 2016 Socio-Economic Survey (SES) conducted by the National Statistical Office (NSO) of Thailand. The SES is a stratified random sample of 43,887 households in Thailand and includes 77 strata, one for each province (Changwat). Each of these strata is separated into two categories as municipal and non-municipal areas.

The SES contains important information on social-economic aspects of the household such as income, expenditure, debts, assets, demographics and characteristics of the dwelling. This study used household data observations in 76 provinces as the unit of analysis. Bangkok Province was excluded because it has a high variation of monthly per capita consumption expenditure. Hence, the observations in this study totaled 41,488.

For out-of-sample prediction, the data was divided into two sets. The initial SES data with 41,488 observations was partitioned into two sub-samples in the ratio 50:50. The first sub-sample or training sample (20,744 observations) was employed to train or fit the model to identify the best model with the optimal set of selected variables. The second sub-sample as the test sample or validation sample (20,744 observations) was used to test the out-of-sample prediction accuracy of the constructed models.

Households from the SES were sampled as a two-stage procedure. Firstly, primary sampling units (PSUs) were randomly selected. Secondly, households within the PSUs were sampled (NSO, 2016). This study randomly sampled PSUs to obtain the training and test samples. Urban households were over-sampled in both the training and test sets since the number of urban household samples in the initial data was greater than the household samples in rural areas. Stepwise regression, LASSO and RF were used to identify and optimize the selection of variables to determine monthly per capita consumption expenditure. The variable selection process was initiated by using stepwise regression, LASSO and RF with 47 variables. After subset selection, the dataset retained only a subset of the selected variables. OLS was used to estimate the coefficients of the retained inputs.

Table 1 shows the numbers of urban and rural household observations in the initial, training and test datasets. In the initial set, the proportions of households living in urban and rural areas were 58.80 and 41.20 percent respectively. The results of data partition both in the training and test datasets provided the proportion of households living in urban and rural areas corresponding to the proportion of the initial set. This data partition was used to estimate the model.

Table 1: Number of Urban and Rural Household Observations of Initial, Training and Test Datasets

	Initial Set		Training Set		Test Set	
	Urban	Rural	Urban	Rural	Urban	Rural
Observations	24,394 (58.80)	17,094 (41.20)	12,226 (58.94)	8,518 (41.06)	12,168 (58.66)	8,576 (41.34)
Total	41,488		20,744		20,744	

Note: Percentage values are in parentheses.
Source: Author's calculation based on SES 2016.

The first step was to identify the variables presented in the SES. These variables were then chosen for the variable selection process in stepwise regression, LASSO and RF approaches to construct the PMT in the OLS model. From the literature, this study considered household assets, demographics, dwelling characteristics and dependency variables that correlated with monthly per capita consumption expenditure. Initially, the model had 47 variables. The variables were classified into five categories consisting of (i) Household characteristics comprised the household head characteristics such as sex, age and marital status. Education level was composed of primary education, lower secondary education, upper secondary education, vocational education and higher education, with also the number of household members and working members. (ii) Dependency comprised household status that had elderly persons aged over 60 years with children aged below 15 and disabled household members. (iii) Housing conditions were characteristics of dwelling and status such as free rent, live with others and dwelling that were constructed by non-permanent or local material such as bamboo. (iv) Ownership of assets. (v) Location as the dummy variable for regions to eradicate differences in the location. However, these dummy variables were not included in the variable selection process (see Table 2).

To consider which condition should be used to screen the poor households, the PMT model estimated monthly per capita consumption expenditure to determine the poor households was compared with the poverty line. Households that had monthly per capita consumption expenditure below the poverty line as 2,667, 2,902 and 2,425 baht per month for national, urban and rural area respectively (NESDB, 2016) were classified as poor household, while households above the poverty line were classified as non-poor household.

Table 2: Variable Description and Basic Statistics

Variable Description	Mean	Std. Dev.	Min	Max
Household Characteristics				
Number of HH members	2.90	1.56	1	14
HHH is female	0.40	0.49	0	1
Age of HHH (Year)	54.01	15.23	12	99
HHH is married	0.65	0.48	0	1
Number of working HH member	1.69	1.05	0	8
HHH with primary education	0.59	0.49	0	1
HHH with lower secondary education	0.10	0.30	0	1
HHH with upper secondary education	0.09	0.28	0	1
HHH with vocational education	0.06	0.24	0	1
HHH with higher education	0.11	0.31	0	1
Dependency				
Proportion of HHM aged < 15 years old	0.13	0.19	0	1
Proportion of HHM aged >= 60 years old	0.25	0.35	0	2
Proportion of HHM disabled	0.04	0.14	0	1
Housing Characteristics				
Number of rooms	2.83	1.23	1	9
Electricity in dwelling	0.99	0.04	0	1
Dwelling constructed with local material	0.004	0.06	0	1
Rental paid by others	0.06	0.24	0	1
Drinking water from the well or underground water	0.05	0.21	0	1
Drinking water from the river, steam, rainwater, etc.	0.13	0.34	0	1
Dwelling has no toilet	0.004	0.07	0	1
Using squat	0.59	0.49	0	1
Ownership of Assets				
Bicycle	0.41	0.49	0	1
Motorcycle	0.82	0.38	0	1
Car	0.17	0.37	0	1
Van or mini truck	0.29	0.45	0	1
Other mini truck	0.11	0.31	0	1
Cooking stove using gas	0.81	0.39	0	1
Cooking stove using electricity	0.15	0.36	0	1
Microwave oven	0.23	0.42	0	1
Electric pot	0.73	0.44	0	1
Refrigerator	0.92	0.28	0	1
Electric iron	0.82	0.38	0	1
Electric cooking pot	0.90	0.31	0	1
Electric fan	0.98	0.14	0	1
Radio	0.44	0.50	0	1
TV	0.77	0.42	0	1
LCD or LED or PLASMA	0.34	0.47	0	1
Video player	0.36	0.48	0	1
Washing machine	0.69	0.46	0	1
Air conditioner	0.24	0.43	0	1
Water boiler	0.20	0.40	0	1
Computer	0.21	0.41	0	1
Telephone	0.06	0.24	0	1
Mobile phone	0.96	0.20	0	1
Fluorescence	0.96	0.20	0	1
Light bulb	0.10	0.30	0	1
Compact fluorescent	0.40	0.49	0	1
Observations	41,488			

Note: HH = household, HHH = household head and HHM = household member.

Source: Author’s calculation based on SES 2016.

4.2 Variable Selection Process

To select the subset of variables that could accurately predict monthly per capita consumption expenditure, the stepwise regression, Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF) were proposed. The training set was used to calibrate a model and then used selected variables from a calibrated model to predict the outcomes (log of monthly per capita consumption expenditure) in the testing set using OLS. Performances of the three algorithms were considered from their predictive accuracy. This was measured by mean squared error (MSE). During the process of calibration, the model was adjusted iteratively to obtain the best performance, while the algorithm identified the subset chosen as the best-selected variables within this process.

Variable selection processes of stepwise regression, LASSO and RF were begun with 47 variables. After the subset selection procedure, the model retained only a subset of variables that stepwise regression, LASSO and RF selected, and the rest was eliminated from the model. OLS was used to estimate the coefficients of the variables that were retained.

4.2.1 Stepwise Regression

To obtain smaller models (smaller selected variables) from a stepwise regression procedure, the results were based on a 0.01 significance level for adding variables to the model and a 0.05 significance level for the removal of variables from the model. The forward stepwise regression was running by the Stata program. This procedure began with an empty model in the training set. If the most-significant removal term is significant, then add it into the model and refit the model. If not, stop. Continue the process. If the least-significant additional term is “insignificant”, then remove it and refit the model. If the most-significant removal term is “significant”, add it and refit the model. Repeat these steps until there is no variable for addition and deletion. The final step involves running OLS to estimate the coefficients of the variables selected by stepwise regression.

4.2.2 LASSO

The variable selection process of Lasso can be defined as the following steps;

(i) Run the LASSO algorithm in the training set using the “*glmnet*” function and assign an alpha value equal to 1, defined as the LASSO function. The training set contains the expected output value (log of monthly per capita consumption expenditure) and 47 candidate variables as an initial step.

(ii) The model that has been trained in a training set is now assessed for accuracy. In this case, using the mean squared error (MSE) as the criterion.

(iii) Tune the model in a validation set to select the optimal lambda value with the lowest mean square error. The LASSO model is trained with k-folds cross-validation, say 10-folds cross-validation using the “*cv.glmnet”* function. Then, the several values of lambda with different numbers of selected variables were obtained. Variables that have coefficients that are not equal to zero are identified as selected variables in this step.

(iv) Perform the model with a selected lambda obtained from the training set to predict the output in a test set as an out-of-sample prediction to evaluate model performance.

(v) Use the OLS method to estimate the coefficients of the LASSO selected variable.

4.2.3 Random Forest

The training set consisted of 20,744 observations with 47 variables (explanatory variables). The Random Forest algorithm was built on multiple models (CART) with different samples and different initial variables. In this case, *n* observations and *m* randomly selected variables were chosen to build the model in two of the three training sets. The remaining training set was left out for constructing the model and called the out-of-bag sample (OOB). This was used to select the variables that provided the lowest OOB error. Therefore, RF used the OOB sample to select the variable that provided the preferred model with the lowest prediction error. Then, it repeated the process (say) 500 times and selected the model that was constructed in a training set to predict the log of monthly per capita consumption expenditure in a test set (out-of-sample) and assess the prediction accuracy of the model.

To clarify the steps of Random Forest, the practical RF working method using the R program and RF packages by Breiman can be described below:

(i) Randomly split the data into two sets as a training set for constructing the model using 20,744 observations and a test set for predicting model performance using 20,744 observations. In this case, the SES dataset has 41,488 households (observations). From 20,744 observations in the training set, the algorithm randomly picks 13,829 observations to construct the model in the training set and the remaining 6,915 observations are used to assess the performance of the model in the OOB procedure to select the number of variables at each split of trees (*mtry*) that provides the lowest MSE value.

(ii) Run the RF algorithm in a training set with the best *mtry* by using the "library(randomForest)" package in R. Then, the RF algorithm creates 500 trees.

(iii) RF has its own variable selection which is called "*Variable Importance*". The variable importance process provides a *%inc mse* value for the regression process. The higher this value, the more importance is assigned to the variable. This has more impact on the dependent variables (log of monthly per capita consumption expenditure).

(iv) Predict and evaluate the accuracy of the model in a test set using the model that was trained in a training set.

(v) Use OLS to estimate the coefficients of the RF selected variable.

4.3 Construction of PMT

Ordinary Least Square (OLS) is the simplest and earliest predictive method for the PMT. A linear combination of independent variables such as household characteristics, household ownership of assets and characteristics of dwelling can be used to estimate a continuous outcome (dependent variable) as the monthly per capita consumption expenditure of the household in terms of the natural logarithm. The objective of the OLS regression model is to estimate the regression coefficient vector β such that the mean squared error (MSE) is minimized.

Given the dataset of *n* household observations, an OLS regression model with *k* explanatory variables can be expressed as:

$$y_i = \alpha_i + \beta_k X_{ik} + \varepsilon_i \, , \; i = 1,2,\ldots,n, \qquad (4.1)$$

where y_i represents the per capita (monthly) consumption expenditure for the i^{th} household, α_i is a constant term, β_k is the coefficient for the k^{th} variable, X_{ik} denotes the set of explanatory variables that are obtained from stepwise regression, LASSO and RF for the k^{th} variable of the i^{th} household and ε_i is the random error term. Then the predicted monthly per capita consumption expenditure can be expressed as:

$$\hat{y}_i = \hat{\alpha}_i + \hat{\beta}_k X_{ik} . \tag{4.2}$$

In practice, the OLS method is used for estimating α_i and β_k implements the log of monthly per capita consumption expenditure as the dependent variable, which can be expressed as:

$$\log(\hat{y}_i) = \hat{\alpha}_i + \hat{\beta}_k X_{ik} . \tag{4.3}$$

The selected variables from the stepwise regression, LASSO and RF that are statistically significant with the log of monthly per capita consumption expenditure following the OLS procedure can be considered as those in the final model. After running the OLS estimation, the coefficients of each variable are used to construct the variable weight. Then, the household is assigned an aggregate score (predicted monthly per capita consumption expenditure of household is also called PMT score) that is a weighted combination of variables. This is calculated as the regression constant plus or minus the weighted variables and each coefficient is multiplied by 100 and rounded to the nearest integer. The equation can be written as:

$$Score_i = \hat{\alpha}_i + \textstyle\sum_{i=1}^{n} x_{ik} \left(\hat{b}_k \times 100\right). \tag{4.4}$$

4.4 Assessing Targeting Accuracy of PMT

The targeting error is adopted to evaluate the targeting accuracy. Grosh and Baker (1995) proposed *Type I* and *Type II* errors to measure inclusion and exclusion errors by categorizing the household into four groups as to whether their true and predicted (by the regression model) monthly per capita consumption expenditure levels fall above or below the cut-off point. From Table 3, households that are likely to be excluded from the beneficial program are classified as *Type I* error cases. By contrast, the households which are incorrectly identified as eligible are classified as a case of *Type II* error. (i) Exclusion error or undercoverage is calculated by dividing the number of *Type I* errors by the total number of households that should obtain the benefit, E_1/N_1. Similarly, (ii) inclusion error or leakage is calculated by dividing the number of *Type II* errors by the number of households that are selected by the program to be beneficiaries as E_2/M_1. Tradeoffs between inclusion and exclusion errors can take place. If the objective is to reduce the budget cost, then decreasing the inclusion errors is preferred. Conversely, if the objective is to increase the welfare of the poor, the alleviation of exclusion error is favored. Poverty accuracy (PA) is the correct prediction of the poor divided by the total of the true poor. This is calculated by

S_1/N_1. Total accuracy (TA) is the sum of the correctly predicted poor and non-poor divided by the total sample. This is calculated by $\frac{S_1+S_2}{S_1+S_2+E_1+E_2}$.

Table 3: *Type I* and *Type II* Errors

	Truly poor $(p=1)$	Non-poor $(p=0)$	Total
Eligible $(\hat{p}=1)$	Targeting success (S_1)	*Type II* error (E_2)	M_1
Ineligible $(\hat{p}=0)$	*Type I* error (E_1)	Targeting success (S_2)	M_2
Total	N_1	N_2	

Source: Author’s summarization.

5. Results and Discussion

5.1 Variable Selection Results and PMT Scores

At the first stage, stepwise regression selected 41, 38 and 38 out of 47 variables for the national, urban and rural levels respectively. To select the same variables with stepwise regression, LASSO selected 41 and 38 variables out of 47 variables for national and rural levels respectively, while the urban model obtained 37 variables since there was no appropriate lambda value to shrink the coefficients to 38 variables. RF also selected 41, 38 and 38 out of 47 variables for the national, urban and rural levels respectively. Then, run the OLS[1] with these variables. After running OLS estimation, the numbers of variables significant with the log of monthly per capita consumption expenditure at 90, 95 and 99 percent confidence levels are shown in final row of Table 4. For the next step, the coefficients of each variable were used to construct the variable weight.

Table 4 shows the weight results from the stepwise regression, LASSO and RF selected variables that are statistically significant with the log of monthly per capita consumption expenditure on OLS procedure. These were considered as the selected variables in the final model. Five variables were not chosen in any model as follows: other mini truck, electric cooking pot, electric fan, TV and fluorescence because almost all households had these items.

Before calculating the PMT performance, rank the actual monthly per capita consumption expenditure of households in descending order in all areas and assigned 2,667, 2,902 and 2,425 baht as the cut-off point of the national, urban and rural areas respectively to classify the poor household. Then, rank the household’s PMT score in descending order and selected PMT scores that had actual monthly per capita consumption expenditure equal to 2,667, 2,902 and 2,425 baht for the national, urban and rural areas respectively. Next, sum their PMT scores to obtain an average score as a cut-off PMT. For example, in the case of the national level, four households had monthly per capita consumption expenditure equaled to 2,667 baht and their PMT scores were 0.42, 0.39, 0.39 and 0.49. The four values of PMT scores were summed and divided by 4; the result was 0.4 as the PMT cut-off score for targeting. After completing this process for nine models, the PMT cut-off score was found to be 0.4 for all models. Any household that had a PMT score of below 0.4 was considered as a poor household in terms of the PMT criteria. Finally, evaluate the targeting accuracy of PMT by calculating the exclusion (*Type I*) and inclusion

[1] Survey regression or svy command was used to run the OLS regression using the Stata program.

(*Type II*) errors. For example, if the household had defined PMT scores higher than 0.4 but the actual monthly per capita consumption expenditure was less than the poverty line, this indicated that they were a truly poor household. The PMT score determined them as a non-poor household which was an exclusion error.

Table 4: Weight Results of the Variables

Variable	Dummy	Weight on each variable								
		I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
Household Characteristics										
Number of HH member		-16.56	-16.64	-16.68	-17.65	-17.11	-18.17	-15.76	-14.85	-15.97
HHH is female	*			-1.74						
Age of HHH (Year)		-0.32	-0.32	-0.36	-0.23	-0.3	-0.29	-0.32	-0.35	-0.38
HHH is married	*	-13.73	-13.61	-14.08	-11.11	-10.34	-10.26	-14.74	-14.25	-14.75
Number of working HH member		2.18	2.19	2.04	2.61		2.54	2.02		1.83
HHH with primary education	*	7.32		-2.81	7.43	-4.74	-7.56	7.69		
HHH with lower secondary	*	16.68	9.85		18.05			15.74	8.02	
HHH with upper secondary	*	20.36	13.48	8.81	23.33	9.67	6.03	17.43	9.42	8.25
HHH with vocational education	*	21.85	14.99	10.11	22.66	9.13		20.44	12.5	11.28
HHH with higher education	*	36.17	29.28	24.27	37.68	23.34	19.29	34.43	26.19	24.58
Dependency										
Proportion of HHM aged < 15 years old		-35.4	-35.31	-35.66	-27.30	-31.21	-26.22	-37.2	-41.82	-37.82
Proportion of HHM aged >= 60 years old		-10.53	-10.41	-9.65	-11.64	-12.6	-11.44	-9.84	-10.53	-8.91
Proportion of HHM is disable		-19.23	-19.26	-19.78	-13.69	-14.43	-14.63	-21.52	-22.62	-22.3
Housing Characteristics										
Number of rooms		2.25	2.26	2.32		1.74	1.73	2.83	2.72	2.70
Electricity in dwelling	*				47.5	47.58				
Dwelling constructed with local material	*	-22.08	-23.28	-25.05				-23.52	-24.72	
Rental paid by others	*					4.57				
Drinking water from the well or underground water	*	-11.57	-11.85	-12.42	-18.36	-18.28	-18.90	-8.22	-8.47	
Drinking water from the river, steam, rainwater, etc.	*	-8.95	-8.92	-9.07	-13.69	-13.45	-13.33	-6.82	-6.52	-6.20
Dwelling has no toilet	*	-23.57	-24.21					-28.19	-30.15	
Using squat	*	-10.7	-10.81	-10.5	-11.13	-11.21	-11.37	-9.61	-9.69	-9.58
Ownership of Assets										
Bicycle	*				-3.26	-3.30	-3.92			
Motorcycle	*	-5.12	-4.73	-4.40	-3.98		-2.88	-6.11	-5.48	-5.17
Car	*	27.57	27.57	27.85	26.84	27.06	26.67	28.97	29.11	29.22
Van or mini truck	*	24.84	24.96	25.07	22.67	23.26	22.99	26.47	26.73	26.87
Other mini truck	*									
Cooking stove using gas	*						-5.15	3.69		
Cooking stove using electricity	*	5.37	5.22	5.16		4.59		6.49	6.19	6.25
Microwave oven	*	8.19	8.18	8.35	9.23	8.57	8.64	8.35	8.15	8.06
Electric pot	*	3.97	4.19	4.66	4.76	4.84	5.23	3.52	3.82	4.37
Refrigerator	*									

Variable	Dummy	Weight on each variable								
		I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
Electric iron	*	11.04	11.44	12.16	11.98	12.43	13.95	9.61	9.83	10.59
Electric cooking pot	*									
Electric fan	*									
Radio	*			-2.22						
TV	*									
LCD or LED or PLASMA	*	6.6	6.63	6.71	6.44	6.64	6.43	6.83	6.7	6.85
Video player	*	6.02	6.04	6.28	4.34	4.25	4.24	7.28	7.17	7.32
Washing machine	*								3.62	3.92
Air conditioner	*	11.93	11.93	11.9	11.88	10.98	10.22	13.00	12.78	13.51
Water boiler	*			2.97			6.52			
Computer	*	8.47	8.47	8.71	8.74	7.89	8.47	9.02	8.95	9.60
Telephone	*	6.61	6.67		9.41	9.53	8.68			
Mobile phone	*	13.65	14.01	14.11	12.48	12.89	14.87	12.68	13.6	13.92
Fluorescence	*									
Light bulb	*	3.65	3.64	4.15	6.27	5.74	5.98			
Compact fluorescent	*	2.53	2.61	2.77				3.27	3.42	3.52
Location										
North	*	-24.4	-24.7	-25.48	-22.43	-24.93	-25.69	-24.52	-25.53	-26.63
Northeast	*	-14.15	-13.74	-14.03	-11.31	-12.74	-12.82	-14.68	-14.48	-15.17
South	*	-4.84	-4.93	-5.44	1.58	-0.47	0.45	-7.99	-7.83	-9.28
Constant	*	891	897	903	848	859	910	886	895	895
R-squared		0.664	0.663	0.661	0.672	0.67	0.67	0.626	0.625	0.62
Observations		20,744	20,744	20,744	12,226	12,226	12,226	8,518	8,518	8,518
N		33	32	33	30	30	30	32	30	27

Note: *I*, *IV* and *VII* are PMT models with stepwise regression selected variables; *II*, *V* and *VIII* are PMT models with LASSO selected variables; *III*, *VI* and *IX* are PMT models with RF selected variables and N is the number of selected variables that are statistically significant with the log of monthly per capita consumption expenditure at 90, 95 and 99 confidence levels respectively.

Source: Author’s calculation.

5.2 Targeting Accuracy of PMT Results

Table 5 to 7 compare the results of selected variables and targeting accuracy performance in each of the nine models. At the national level (Table 5), PMT with the selected variables from stepwise regression (*Model I*) captured the highest total number of poor and non-poor households (89.11 percent of total accuracy) and provided the lowest number of non-poor households that were classified as poor households (46.00 percent of an inclusion error). For the poverty accuracy and exclusion error, the PMT model with selected variables of RF (*Model III*) was the best model at capturing the actual poor households (73.83 percent of the poverty accuracy) and reduced the number of actual poor households that were classified as non-poor households (26.17 percent of the exclusion error). When comparing the PMT models by the selected variables from the machine learning approach with the traditional PMT model (selected variables from a stepwise regression), results indicated that PMTs with the selected variables from the LASSO (*Model II*) and RF (*Model III*) outperformed the traditional PMT (*Model I*) in terms of increase in poverty accuracy and reduction in exclusion error.

Table 5: Targeting Performance of PMTs at the National Level

Variable	National		
	I	*II*	*III*
Total Accuracy (TA)	18,484 (89.11)	18,425 (88.82)	17,548 (84.59)
Poverty Accuracy (PA)	1,228 (50.29)	1,268 (51.92)	1,803 (73.83)
Inclusion Error (IE)	1,046 (46.00)	1,145 (47.45)	2,557 (58.65)
Exclusion Error (EE)	1,214 (49.71)	1,174 (48.08)	639 (26.17)
PMT cut-off score	0.4	0.4	0.4
Poverty line (Baht)	2,667	2,667	2,667
Poor households in SES	2,442	2,442	2,442
Observations	20,744	20,744	20,744

Note: Percentage values are in parentheses with the number of households classified as poor and non-poor above.
Source: Author’s calculation.

At the urban level (Table 6), PMT with the selected variable from stepwise regression (*Model IV*) captured the highest total number of poor and non-poor households (85.50 percent of total accuracy) and provided the lowest number of non-poor households that were classified as poor households (56.26 percent of an inclusion error). For poverty accuracy and exclusion error, the PMT model with selected variables of RF (*Model VI*) was the best at capturing the actual poor households (87.16 percent of the poverty accuracy) and reduced the number of actual poor households that were classified as non-poor households (12.84 percent of the exclusion error). When comparing the PMT models by the selected variables from machine learning approach with the traditional PMT model (selected variables from a stepwise regression), results indicated that PMTs with the selected variables from the LASSO (*Model V*) and RF (*Model VI*) outperformed the traditional PMT (*Model IV*) in terms of increase in poverty accuracy and reduction in exclusion error.

Table 6: Targeting Performance of PMTs at the Urban Level

Variable	Urban		
	IV	*V*	*VI*
Total Accuracy (TA)	10,404 (85.50)	10,333 (84.92)	8,886 (73.03)
Poverty Accuracy (PA)	1,076 (73.90)	1,116 (76.65)	1,269 (87.16)
Inclusion Error (IE)	1,384 (56.26)	1,495 (57.26)	3,095 (70.92)
Exclusion Error (EE)	380 (26.10)	340 (23.35)	187 (12.84)
PMT cut-off score	0.4	0.4	0.4
Poverty line (Baht)	2,902	2,902	2,902
Poor households in SES	1,456	1,456	1,456
Observations	12,168	12,168	12,168

Note: Percentage values are in parentheses with the number of households classified as poor and non-poor above.
Source: Author's calculation.

At the rural level (Table 7), PMT with the selected variable from the LASSO (*Model VIII*) captured the highest total number of poor and non-poor households (88.14 percent of total accuracy) and provided the lowest number of non-poor households that were classified as poor households (53.31 percent of an inclusion error). For poverty accuracy and exclusion error, the PMT model with selected variables of RF (*Model IX*) was the best at capturing the actual poor households (79.75 percent of the poverty accuracy) and reduced the number of actual poor households that were classified as non-poor households (20.25 percent of the exclusion error). When comparing the PMT models by the selected variables from the machine learning approach with the traditional PMT model (selected variables from a stepwise regression), results indicated that PMTs with the selected variables from the LASSO (*Model VIII*) and RF (*Model IX*) outperformed the traditional PMT (*Model VII*) in terms of increase in poverty accuracy and reduction in exclusion error.

Table 7: Targeting Performance of PMTs at the Rural Level

Variable	Rural		
	VII	*VIII*	*IX*
Total Accuracy (TA)	7,525 (87.86)	7,559 (88.14)	6,836 (79.71)
Poverty Accuracy (PA)	505 (52.99)	452 (47.43)	760 (79.75)
Inclusion Error (IE)	593 (54.01)	516 (53.31)	1,547 (67.06)
Exclusion Error (EE)	448 (47.01)	501 (52.57)	193 (20.25)
PMT cut-off score	0.4	0.4	0.4
Poverty line (Baht)	2,425	2,425	2,425
Poor households in SES	953	953	953
Observations	8,576	8,576	8,576

Note: Percentage values are in parentheses with the number of households classified as poor and non-poor above.
Source: Author's calculation.

Therefore, PMTs with the selected variables using the machine learning approach accurately targeted the number of actual poor households at all levels. PMTs with selected variables of stepwise regression outperformed all models in case of overall targeting accuracy rate and reduced inclusion error at the national level only. Moreover, the results showed a trade-off between two errors as decreasing the exclusion error tended to increase the inclusion error. Results suggested that using RF in terms of variable selection for constructing PMTs provided the best results of reducing exclusion error and increasing poverty targeting, better than traditional PMTs with stepwise regression selected variables. However, the PMT using selected variables of RF had the lowest total accuracy at all levels. This finding was supported by McBride and Nichols (2016) who found that the quantile RF did not improve the total accuracy in all countries (Bolivia, Malawi and East-Timor Leste).

Then, we have considered the variables in the model that explained the monthly per capita consumption expenditure with significance at 90, 95 and 99 percent confidence levels in Table 8. Overall, 20 variables were selected in all models, consisting of the Number of HH members, Age of HHH (Year), HHH is married, HHH with upper secondary education, HHH with higher education, Proportion of HHM aged < 15 years old, Proportion of HHM aged >= 60 years old, Proportion of HHM is disabled, Drinking water from the river, Using squat, Car, Van or mini truck, Microwave oven, Electric pot, Electric iron, LCD or LED or PLASMA, Video player, Air conditioner, Computer, and Mobile phone. Most of these selected variables were consistent with a study by Punyasavatsut (2017) who assessed dependency members, housing conditions and ownership of assets.

At the national level, *Model III* (RF selected variables) was the best at targeting the actual poor households and provided the different variables from *Models I* and *II* as household head is female, radio and water boiler. This implied that poor households tended to have a female as the household head (with monthly per capita consumption expenditure less than the male as a household head), while households that had a radio and water boiler were likely to be poor households. At the urban level, *Model VI* (RF selected variables) provided different variables from *Models IV* and *V* as ownership of cooking stove and using gas and water boiler. This implied that households that used a cooking stove with gas and water boiler are likely to be poor households in an urban area. In the rural area, *Model IX* (RF selected variables) provided a set of variables similar to *Models VII* and *VIII*. However, *Model IX* provided a number of selected variables that were significantly less than *Models VII* and *VIII*, even though the variable selection began with the same number. The variables that appeared in *Models VII* and *VIII* but were excluded from *Model IX* consisted of HHH with lower secondary education, the dwelling constructed with local materials, drinking water from the well or underground water and dwelling has no toilet. This indicated that these variables are not good enough to target poor households in rural areas.

Table 8: Selected Variables in each Model

	National			Urban			Rural		
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
Household Characteristics									
Number of HH members	x	x	x	x	x	x	x	x	x
HHH is female			x						
Age of HHH (Year)	x	x	x	x	x	x	x	x	x
HHH is married	x	x	x	x	x	x	x	x	x
Number of working HH members	x	x	x	x		x	x		x
HHH with primary education	x		x	x	x	x	x		
HHH with lower secondary education	x	x		x			x	x	
HHH with upper secondary education	x	x	x	x	x	x	x	x	x
HHH with vocational education	x	x	x	x	x		x	x	x
HHH with higher education	x	x	x	x	x	x	x	x	x
Dependency									
Proportion of HHM aged < 15 years old	x	x	x	x	x	x	x	x	x
Proportion of HHM aged >= 60 years old	x	x	x	x	x	x	x	x	x
Proportion of HHM disabled	x	x	x	x	x	x	x	x	x
Housing Characteristics									

	National			Urban			Rural		
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*
Number of rooms	x	x	x		x	x	x	x	x
Electricity in dwelling				x	x				
Dwelling constructed with local material	x	x	x				x	x	
Rental paid by others					x				
Drinking water from the well or underground water	x	x	x	x	x	x	x	x	
Drinking water from the river, steam, rainwater, etc.	x	x	x	x	x	x	x	x	x
Dwelling has no toilet	x	x					x	x	
Using squat	x	x	x	x	x	x	x	x	x
Ownership of Assets									
Bicycle				x	x	x			
Motorcycle	x	x	x	x		x	x	x	x
Car	x	x	x	x	x	x	x	x	x
Van or mini truck	x	x	x	x	x	x	x	x	x
Other mini truck									
Cooking stove using gas						x	x		
Cooking stove using electricity	x	x	x		x		x	x	x
Microwave oven	x	x	x	x	x	x	x	x	x
Electric pot	x	x	x	x	x	x	x	x	x
Refrigerator									
Electric iron	x	x	x	x	x	x	x	x	x
Electric cooking pot									
Electric fan									
Radio			x						
TV									
LCD or LED or PLASMA	x	x	x	x	x	x	x	x	x
Video player	x	x	x	x	x	x	x	x	x
Washing machine								x	x
Air conditioner	x	x	x	x	x	x	x	x	x
Water boiler			x			x			
Computer	x	x	x	x	x	x	x	x	x
Telephone	x	x		x	x	x			
Mobile phone	x	x	x	x	x	x	x	x	x
Fluorescence									
Light bulb	x	x	x	x	x	x			
Compact fluorescent	x	x	x				x	x	x
Number of variables are used in the models	33	32	33	30	30	30	32	30	27

Source: Author's calculation.

6. Conclusions

Based on Thailand's Socio-Economic survey data (SES) in 2016, this study stipulated and evaluated a series of multiple regression-based PMTs in terms of model performance in out-of-sample prediction and targeting accuracy performance. LASSO provided signs of explanatory variables that were relative to the monthly per capita consumption expenditure; however, statistical significance of the set of variables was not observed. Similarly, RF was unable to interpret the set of variables obtained from this model since the explanatory variable coefficients did not exist. Variable selection of RF can only determine which variable has the most influence on the dependent variable; known as "Variable Importance". The best variables for prediction can be selected but it was not possible to interpret how these variables affected the dependent variable. Therefore, a two-step procedure was proposed to solve the problem of variable interpretation. First, OLS regression was run with only a set of selected variables from the LASSO and RF algorithms. Then, the variables that were significant with the log of monthly per capita consumption expenditure were selected to construct the PMT scores. The targeting accuracy performance revealed that traditional PMTs, based on variable selection of stepwise regression, performed better for total accuracy and inclusion error than PMTs with LASSO and RF at the national and urban levels. At the rural level, the PMT with LASSO performed better. On the other hand, PMTs based on the variable selection of RF performed better in terms of poverty accuracy and exclusion error at all levels.

For policy implication, implementation of the PMT for targeting poor households should recognize the different areas of poverty. This study also suggests that PMTs based on the variable selection of stepwise regression are more appropriate in terms of overall targeting accuracy. These can target the actual poor and non-poor households and reduce inclusion error better than PMTs based on selected variables from LASSO and RF. The PMTs with the selected variables of RF reduced an exclusion error more than PMTs with the selected variables of stepwise regression and LASSO. However, when comparing the performance of PMTs based on LASSO and RF, the results were confronted with a trade-off between exclusion and inclusion errors. If policymakers are concerned about the budget burden of the program, then PMTs based on the variable selection of LASSO are the suitable choices at the rural level, while PMTs with variable selection of stepwise regression are appropriate choices for national and urban levels. By contrast, if the policymakers would like to cover poor households, PMTs based on the variable selection of RF are more appropriate because this can reduce exclusion errors and accurately target poor households.In the real world, policymakers are faced with inclusion and exclusion errors that can cause social contradictions based on predictive models. Policymakers should define the acceptance level of exclusion errors and set the poor design or cut-off points at appropriate levels.

Finally, this study had the following data limitations. The SES data used here were in a one-year cross-sectional form. Results may not be consistent in the future because if the model is trained in another dataset, the predictive result can change since the nature of next year’s dataset will not concur with previous years. In a future study, the author would like to improve the PMT model based on the LASSO and RF algorithms by using two years of SES data to examine the overtime results of poor household targeting.

Acknowledgement

I would like to express my deepest gratitude to Assoc. Prof. Dr. Chaiyuth Punyasavatsut, my adviser, who expanded my vision and knowledge of the PMT to target poor households, Assoc. Prof. Dr. Tanapong Potipiti who offered pertinent advice concerning the LASSO and RF algorithms and Asst. Prof. Dr. Chayanee Chawanote who provided meticulous advice from the beginning through to the completion of this article.

References

Ahmed, A. U., & Bouis, H. E. (2002). Weighing what’s practical: proxy means tests for targeting food subsidies in Egypt. *Food Policy*, *27*(5), 519-540. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0306919202000647

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7), 1545-1588. Retrieved from https://doi.org/10.1162/neco.1997.9.7.1545

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521-547. Retrieved from https://projecteuclid.org/euclid.bj/1363192037

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140. Retrieved from https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. Retrieved from https://doi.org/10.1023/A:1010933404324

Brown, C., Ravallion, M., & Van De Walle, D. (2016). *A poor means test? econometric targeting in Africa.* The World Bank. Retrieved from http://hdl.handle.net/10986/25814

Chan-Lau, J. (2017). Lasso Regressions and Forecasting Models in Applied Stress Testing. *IMF Working Papers* No.17/108. Retrieved from http://dx.doi.org/10.5089 /9781475599022.001

Chanmorchan, P., Pornwalai, T., & Popivanova, C. (n.d.). Thailand's child grant support programme. Retrieved from https://transfer.cpc.unc.edu/wp-content/uploads /2016/04/18-Thailands-Child-Grant-Programme.pdf

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1-22. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/

Grosh, M., & Baker, J. L. (1995). *Proxy means tests for targeting social programs: simulations and speculation (English)*. Living Standards Measurement Study Working Paper No. 118. The World Bank. Retrieved from http://documents .worldbank.org/curated/en/750401468776352539/Proxy-means-tests-for-targeting-social-programs-simulations-and-speculation

Tibshirani, R., Wainwright, M., & Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). Overview of supervised learning. ~~In~~ *The elements of statistical learning* (pp. 9-41). Springer, New York, NY. Retrieved from https://doi.org/10.1007/978-0-387-84858-7

Ho, T. K. (1995, August). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition*, 1, 278-282. IEEE. Retrieved from https://ieeexplore.ieee.org/document/598994

Houssou, N., Zeller, M., Alcaraz V, G., Schwarze, S., & Johannsen, J. (2007). *Proxy Means Tests for Targeting the Poorest Households -- Applications to Uganda*. Paper presented at the European Association of Agricultural Economists. Retrieved from http://ageconsearch.umn.edu/record/7946/files/sp07ho01.pdf

James, F. C., & McCulloch, C. E. (1990). Multivariate analysis in ecology and systematics: panacea or Pandora's box?. *Annual review of Ecology and Systematics*, *21*(1), 129-166. Retrieved from https://doi.org/10.1146/annurev.es.21.110190.001021

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491-95. Retrieved from http://dx.doi.org/10.1257/aer.p20151023

Knippenberg, E., Jensen, N., & Constas, M. (2017). Resilience, Shocks, and the Dynamics of Food Insecurity Evidence from Malawi. Retrieved from https://pdfs.semanticscholar.org/6fac/f0b44239fd283e98b9645c9c2127e2d46933.pdf

Koenker, R. W., & Bassett, G. (1978). Regression Quantiles. *Econometrica,* 46(1), 33-50. Retrieved from https://EconPapers.repec.org/RePEc:ecm:emetrp:v:46:y:1978 :i:1:p:33-50

Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. Retrieved from arXiv preprint arXiv:1711.06813.

McBride, L., & Nichols, A. (2015). Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools. Retrieved from http://www.econthatmatters.com/wp-content/uploads/2015/01/improvedtargeting_21jan2015.pdf

McBride, L., & Nichols, A. (2016). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review, 32*(3), 531-550. Retrieved from http://documents.worldbank.org/curated/en/352211475589592980/Retooling-poverty-targeting-using-out-of-sample-validation-and-machine-learning

Narayan, A., & Yoshida, N. (2005). Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka. *Report* No. SASPR–7*, Washington, DC: World Bank,* Retrieved from http://documents.worldbank.org/curated/en/803791468303267323/pdf/332580PAPER0SASPR17.pdf

National Economic and Social Development Board. (n.d.). Report of Poverty and Inequality Circumstance in Thailand 2016. Retrieved from https://transfer.cpc.unc.edu/wp-content/uploads/2016/04/18-Thailands-Child-Grant-Programme.pdf

Nguyen, C., & Lo, D. (2016). Testing Proxy Means Tests in the Field: Evidence from Vietnam. Retrieved from https://mpra.ub.uni-muenchen.de/id/eprint/80002

National Statistical Office. (n.d.). The 2016 Household Social-Economics Survey in Thailand. Retrieved from http://ddi.nso.go.th/index.php/catalog/220

Otok, B. W., & Seftiana, D. (2014). The classification of poor households in Jombang With random forest classification and regression trees (RF-CART) approach as the solution in achieving the 2015 Indonesian MDGs' targets. *International Journal of Science and Research (IJSR) Volume 3*. Retrieved from https://www.ijsr.net/archive/v3i8/MDIwMTU1NDA=.pdf

Punyasavatsut, C. (2017). The development of information system for learning opportunities insurance (in Thai). *Economic Research and Training Center (ERTC)*, Faculty of Economics, Thammasat University.

Ravallion, M. (1999). *Issues in measuring and modeling poverty*. The World Bank. Retrieved from http://documents.worldbank.org/curated/en/965061468739145705/Issues-in-measuring-and-modeling-poverty

Sohnesen, T. P., & Stender, N. (2017). Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment. *Poverty & Public Policy, 9*(1), 118-133. Retrieved from https://doi.org/10.1002/pop4.169

Thoplan, R. (2014). Random forests for poverty classification. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, 17(2), 252-259. Retrieved from https://pdfs.semanticscholar.org/370a/5c135812f4a13438eab6fd379de02f929339.pdf

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288. Retrieved from https://www.jstor.org/stable/2346178