Reverse Regression to Understand Unconditional Cash Transfers in Kenya

MengChen Chung
03/19/2021

Summary

The goal of this project is to leverage the tools of machine learning (ML) to build a "discovery engine". We use data from a large-scale randomized controlled trial (RCT) designed to test the effects of receiving an unconditional cash transfer. By building machine learning models to predict who was assigned to which condition, we can expect to learn which outcomes are associated with receiving the cash transfer. The aim is not to build the best predictive model, rather it is to discover outcomes that may have been missed by a traditional theory-driven approach. Through different models and datasets with different variable combinations, we demonstrate the potential to predict randomness. By cross-comparing the important variables in our models, we are able to excavate insightful variables that were effectively "left on the table" in previous research. We also incorporate unsupervised learning approaches to spot the underlying structures of the dataset, showcasing the isolation of pure control observations and similarity between treatment and spillover observations.

Introduction

President Biden recently signed a $1.9 trillion dollar stimulus package in the United States. Though slightly smaller than the $2.2 trillion dollar package the year before (the CARES Act), this bill includes among its benefits a number of new features that has led the New York Times to call it "the biggest antipoverty effort in a generation". Most notably, the Child Tax Credit will provide direct monthly payments to families making less than $150k per year which will affect more than 93% of children in the United States. While on the surface the notion of getting money to those who need it seems like a good thing, skeptics of this largely partisan bill believe that it is too expensive, will not have

the intended effects, and may in fact create negative long term outcomes. This is the struggle of policy makers.

Social scientists have long tried to answer questions like these, but their experiments have been forever plagued by two large limitations, control and scale. While randomized controlled trials (RCTs) have become the gold standard in the field, the standard econometric framework penalizes testing too many outcome measures. If one wanted to test for multiple hypotheses they would need to gather more observations. Researchers, therefore, invest time, money, and effort collecting hundreds, sometimes thousands, of metrics only to later restrict their primary analyses to a handful of outcomes. Even then, which outcomes should a researcher look at? In the case of cash transfers one obvious outcome is to look at how the money is used. Is it saved or spent? If spent, spent on what? If saved, does the amount saved matter? Not to mention that cash transfers could affect other things, such as access to healthcare, employment, stress, happiness, and social activities people engage in. In the end, only a few of these measures can be tested with confidence.

Machine learning offers a solution (Ludwig, Mullainathan, & Spiess, 2019). By using the random treatment assignment as a label to be predicted and deploying machine learning methods to classify the condition participants were assigned to we are able to learn associations between outcome variables and treatment assignment.

Here we used data from a large RCT conducted in Kenya between 2011 and 2013. We fit a series of machine learning models to predict which treatment participants were assigned to. By comparing the variable importance across models, we get an indication of which measures are most

predictive of treatment assignment. In the end we are able to build a set of models which accurately classify treatment assigned and discover relationships not previously discussed in the original research.

The experiment & the data

From 2011 to 2013, Johannes Haushofer and Jeremy Shapiro led an RCT to better understand the effects of unconditional cash transfers. The researchers worked with the NGO GiveDirectly to issue cash transfers to poor households in Kenya. The study surveyed 2,880 individuals, and evaluated over 981 variables for these individuals. The data is shallow and wide.

The goal of this study was to better understand the short and long term impact of unconditional cash transfers (UCT) on poor communities. On average, the transfers in this study were $709 PPP which amounts to roughly 2 months worth of expenditures for families in these villages. In order to glean as much insight as possible from this study, Haushofer and Shapiro set up several layers of randomization for this study. They began at the village level with treatment and control villages. They then increased the granularity by adding a household layer within the treatment villages. Even within the treatment households, the team decided to vary who within the household received the transfer, the matriarch or patriarch. This led the team to define three treatment levels since outcomes might be affected by treatments on one's neighbors:

1. Treatment = In a treatment village, and their household received the treatment
2. Spillover = In a treatment village, but their household did not receive the treatment
3. Pure Control = Not in a treatment village

The researchers looked at a huge variety of outcomes from their treatment. As stated earlier, there are 981 columns in the data set. For this analysis, we investigated multiple versions of the data, discussed

later in the Pre-Processing section, in order to discover what may have been missed by the original researchers.

Application of Machine Learning

Machine learning offers a set of tools for making predictions. Whether that is predicting the fastest route home or when an airplane engine will be in need of repair, machine learning looks for patterns in existing data to predict future cases.

For this project, we repurpose machine learning techniques to create a "discovery engine" (Ludwig, Mullainathan, & Spiess, 2019). We use the dataset collected as part of a large-scale randomized controlled trial (RCT) and machine learning tools to predict treatment assignment. In doing so, we explore the relationship between various outcome measures and the treatment assignment.

The intuition behind this approach is that of a reverse regression, whereby the treatment assignment, a predictor in hypotheses tests, is now used as the label we aim to predict. This approach allows us to reach beyond the initial set of hypotheses, by exploring which outcome variables are most important in separating the treatment groups.

Traditional approaches rely on the randomized treatment assignment to draw causal conclusions between receiving the treatment and a single outcome variable. Notice, however, that the randomized treatment assignment serves a very different purpose when used as part of a discovery engine. Here, we are looking to predict who received which treatment; yet, assuming the randomization was done correctly, it is unlikely that our prediction will do significantly better than chance. That is, *if* receiving the treatment had no effect on any of the outcome metrics. If, instead, the model classifies the randomized treatment assignment better than chance this would suggest that something about the

outcome variables is associated with the different conditions. In other words, machine learning allows us to capture the signature patterns across the outcome variables that best reflects each treatment.

Pre-Processing

Prior to fitting the data we cut the data up in three different ways. One included all the original variables (data.wide).[1] The other maintained only the indices created by the original investigators and a set of household characteristics (data.slim). The third removed the data indices created and maintained only endline variables (data.discover).

Data.wide - In order to put the machine learning models to the test we wanted to see what we would learn if all the available data were used to fit the model. We anticipated that this data would contain the most signal, but also the most noise. Many of the variables were derived from others; for example, the indices were a combination of other variables in the dataset. There were also a number of indicator variables which the researchers maintained to indicate missing values, say if a participant did not respond to a survey question. Finally, this dataset included all of the responses collected at the end of the study ('endline') as well as all of the responses collected at the beginning, before any money was dispersed ('baseline'). The hundreds of columns also made it computationally expensive to fit more models using more complex approaches. For these reasons we created two more narrower datasets.

Data.slim - We decided to restrict the columns to include only the aggregated variables, or indices, which the researchers constructed. This is effectively the pre-processed part of data. We therefore dropped any columns that did not contain the terms "index" or "total" and from those we

[1] There were a number of dummy coded variables which indicated the assigned treatment, these were removed in all the cuts.

dropped any that ended with "0" (as these indicated a baseline measure). While this significantly reduced the width of the dataset, from over 900 to under 50, relying on the indices would make it challenging to effectively "discover" any new relationships. The pre-preprocessing reduced the dimensionality of the data but it also masked what exactly within the index was driving the effect. The indices also represented a set of variables which the researchers hypothesized would be affected by the cash transfer. In other words, for the purpose of discovery, we wanted to focus our attention on measures the researchers may not have had the statistical power to explore.

Data.discover - Here we again removed the columns associated with baseline responses but this time we also removed the indices, maintaining instead the variables that *did not* contain "index" or "total". In doing so we were left with approximately 350 columns.

Due to the limits of our computational power, our applications and discussion will focus mostly on *data.slim* and *data.discover*.

Unsupervised Learning

The wide data contains the most comprehensive information, while it also suffers from collinearity. For analysis and interpretability, we need to use other cleaner datasets. The slim data contains 70 variables, and nearly all of them are highly correlated to cash transfer based on domain knowledge, our exploration began in this dataset. First of all, we would like to harness unsupervised learning methodology to explore underlying patterns to reinforce our hypothesis. That is, we would like to examine if we could predict randomness or discern interesting phenomena from an unlabeled approach in this dataset. Therefore, we introduced UMAP to observe the global and local structures on a 2D plane. UMAP identified the underlying relationship between observations, and we colored each

of them with the true labels (**Figure 1**). Interestingly, we found out there were two clear clusters in the original space - the pure control group, and the others. The plots indicate the dissimilarity between the pure control group and the other two classes, and also the similarity between the treatment and the spillover participants.

To be more rigorous, we employed a self-organizing map (SOM), which is also an unsupervised learning procedure to capture global and local structures in high dimensions then project to a low dimensional space (**Figure 2,** red: PureControl, blue: Spillover, green: Treatment). Again, we could see the pure control observations were similar to each other and formed a group, and the other two labels mingling together. We then overlapped k-means and FCM clustering algorithms on the self-organizing map (**Figure 3**). The outcomes demonstrated that these two clustering methods could also spot the pure control to be a group, and the other two classes share large similarities. We then repeated these steps on data.discover (**Figure 4, 5, 6**). Noticeably, UMAP and FCM were not as informative as them on data.slim, probably due to noises, while SOM and k-means still illustrated the pure control group pattern and the mix in treatment and spillover observations.

These findings from the unsupervised perspective reinforced the fact that randomness is predictable in this dataset. Moreover, those who did not receive cash transfer behaved similarly to the treatment group. A possible explanation is regarding network effect. Since spillover was basically the neighbors of the treatment group, they might have enjoyed some “spillovers” from the treatment participants. For instance, the people who received cash transfers were able to buy new equipment on the farm so neighbors became more productive too. Or those who received cash transfers were less stressed then there was less disagreement between them and their neighbors, so everyone was less

stressed. In effect, the original paper mentioned that people responded with similar reasons for being less stressed in the survey. To sum, the pattern from unsupervised learning suggests new insight between treatment and spillover groups, and this pattern can be further tested and be informative in other experiments.

Multiple Models Predicting Randomness

With the interesting underlying data structures in mind, we employed supervised learning methods to further analyze the dataset. We split data into 80/20 for analysis, also ensuring the stratification and distribution in the two datasets are similar to prevent bias. The first model we chose logistic regression, with lasso we could discern important variables, which cannot be easily extracted from theory-based approaches. We then incorporated random forest and boosted tree methods since they tend to provide better prediction accuracy when the assumptions in logistic regression cannot hold. Furthermore, we harnessed less interpretable ML approaches - SVM and neural network - to observe if the prediction can stand in generalized situations. The results showcased that logistic regression with lasso returned prediction accuracy 0.770, random forest and boosting returned 0.943 and 0.953 respectively, SVM with linear kernel had 0.756 accuracy, SVM with radial kernel increased to 0.860, and neural network received 0.777. Overall, all models attained over 0.75 prediction accuracy, indicating the ability to predict randomness.

However, the prediction accuracy from the slim data seemed flutatuating, intimating some important variables might not have been included. One possibility is that those variables grouped in the original paper masked other interesting and important variables, and therefore, we used data.discover, which is the same as data.wide without indices variables (groups created in the original paper) and

baseline variables. For computational feasibility, we did not incorporate logistic regression with lasso, and we observed that random forest returned 0.941, boosting returned 0.950, SVM with linear kernel returned 0.817, SVM with radial kernel returned 0.946, and neural network reported 0.887 in accuracy. Obviously, the forecasting performance did not only improve but also became more stable, which supports our hypothesis. And again, the information from the variables projected randomness.

Variable Importance Across Models

We then shift our focus to interesting variables, aiming to explore factors that had not been discovered in the original paper. We extracted important variables from logistic regression with lasso, random forest, boosted tree, and implemented variable selection with the "Boruta" package in both data.slim and data.discover (data.discover did not contain lasso information), see **Exhibits 2 & 3**. Then we spotted some important variables which are cross-listed in different models.

The original researchers had eight primary conclusions:

1. Transfers allow poor households to build assets.
2. Transfers increase consumption.
3. Transfers reduce hunger.
4. Transfers do not increase spending on alcohol and tobacco.
5. Transfers increase investment in and revenue from livestock and small businesses.
6. Transfers increase psychological well-being of recipients and their families.
7. Transfers affect many, but not all, indicators of poverty.
8. Specific design features of cash transfer programs differentially affect impacts and imply policy trade-offs.

In our modeling, we found that indeed total assets (asset_total_ppp1) and total consumption (cons_total_ppp1) were consistently highly important variables when predicting the treatment. This gave us confidence that our models were at least intuitively in line with the original researchers.

We did discover a few very important variables that the initial researchers seemed to have missed. In particular when modeling with the slim data we found that the members of the household (hh_totalmembers0 and hh_children0) was a very important predictor of the treatment. Though the implication of this finding is not immediately apparent, one hypothesis is that the marginal difference for a small household is much larger than that for a large household. No matter the cause, this finding definitely implies that this attribute needs to be scrutinized in more detail and considered by policy makers proposing UCTs. Similarly, we found that income generating activities (hh_workactivites) was also an important variable for our models. Understanding how this might be affected is *critical* for the long term efficacy of such a policy because it affects employment.

When we took out the indexes and interrogated the raw variables, a few more interesting factors floated to the top of the importance list. Total assets and total consumption were still large factors, but particularly in the random forest, we found that many psychological outcomes related to women, depression (b_f_cesd12), anxiety (b_f_scheier3) and self-esteem (b_f_rosenberg4_surrogate) were also important variables. This reaffirms the researchers' original findings, but also provides the additional insight of their gender bias toward women.

Conclusion

Randomized controlled trials are considered to be the gold standard approach in policy evaluation. However, given the cost associated with implementation of the interventions and data

collection, the datasets are often wide but short. The econometric framework that is traditionally deployed requires that some outcome measures are prioritized (based on theory, expertise, and intuition) while others are cast aside.

Machine learning tools can be used to more fully explore the data and effectively discover relationships that may otherwise have been missed. We demonstrate how this could be done by using data from a randomized controlled trial that examined the effects of receiving an unconditional cash transfer. We fit models to classify which treatment each observation belonged to. In doing so, we were able to flexibly fit models that used all the outcome variables available. Our models were able to predict treatment assignments better than chance. The most important variables are candidate features to be studied in future research.

Appendix

Figure 1. UMAP in data.slim

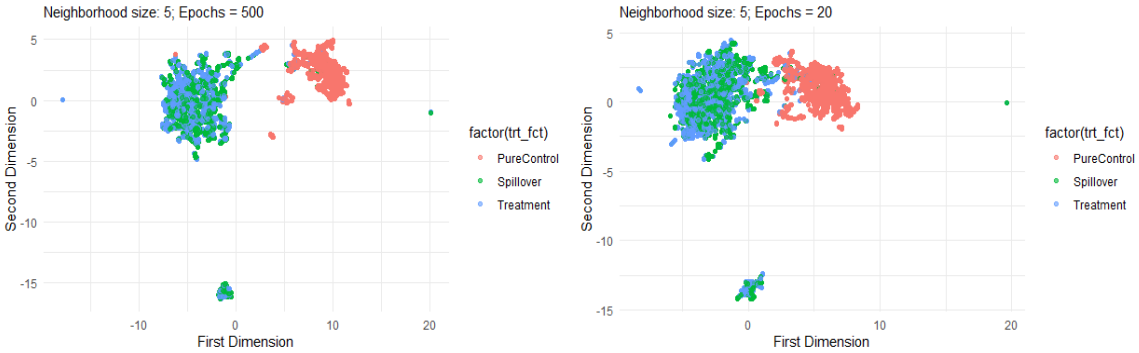

Figure 2. SOM in data.slim

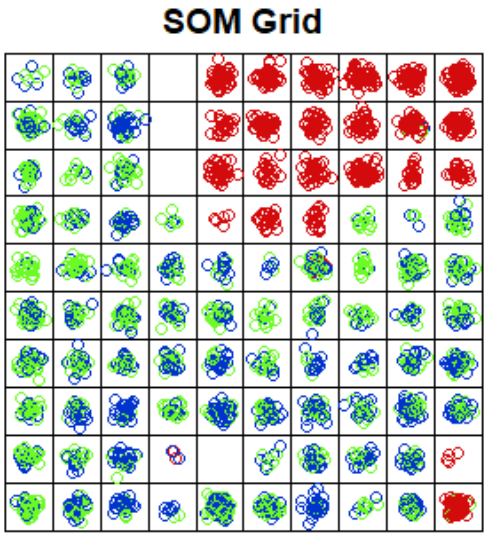

Figure 3. SOM with k-means and FCM in data.slim

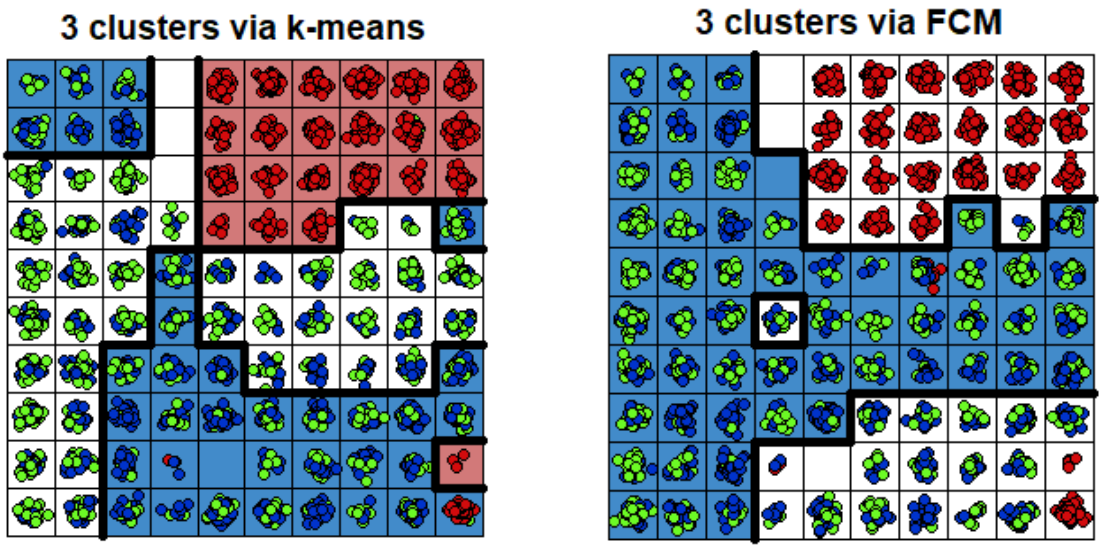

Figure 4. UMAP in data.discover

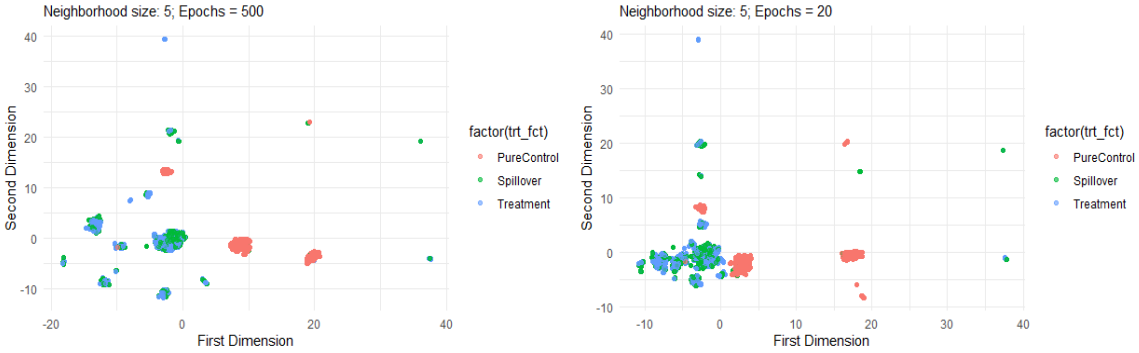

Figure 5. SOM in data.discover

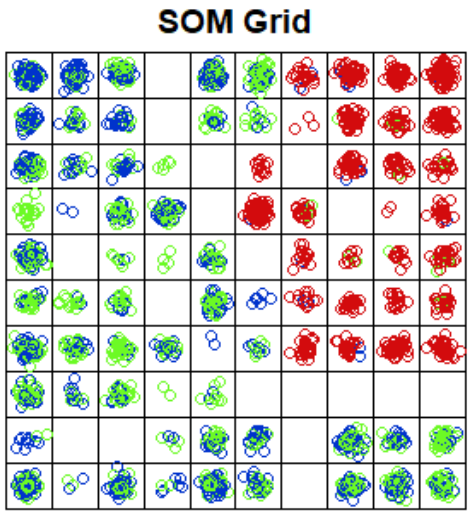

Figure 6. SOM with k-means and FCM in data.slim

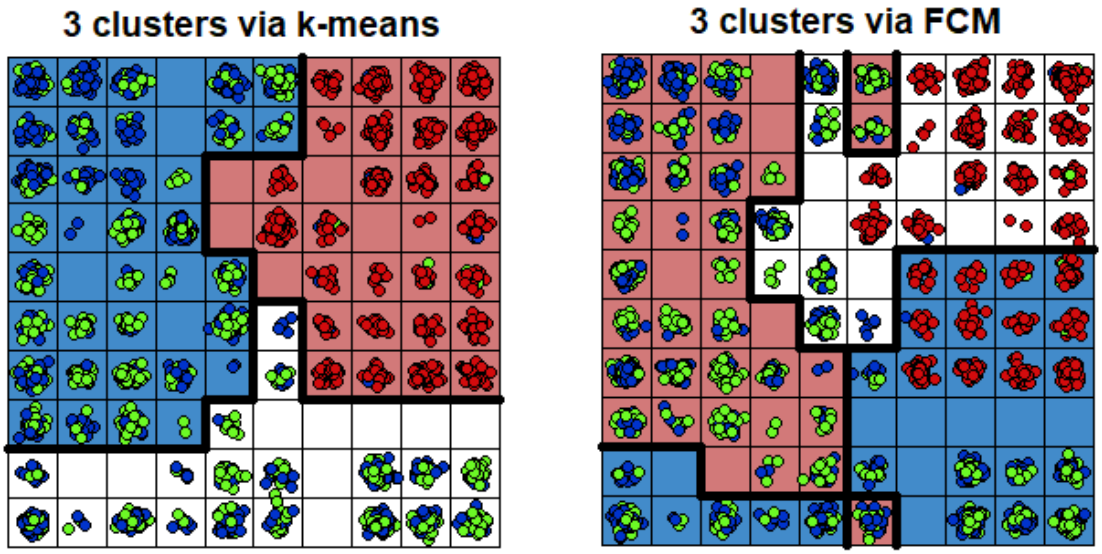

Exhibit 1. Variable Index

Variable	Description
$fs_hhfoodindexnew	"Food security index"
$med_hh_healthindex	"Health index"
$ed_index	"Education index"
$psy_index_z	"Psychological well-being index"
$ih_overall_index_z	"Female empowerment index"
$fs_childfoodindexnew	"Food security index (children)"
$med_child_healthindex	"Health index (children)"
$hh_totalmembers	"Household size"
$asset_total_ppp	"Value of non-land assets (USD)"
$ent_total_rev_ppp	"Total revenue, monthly (USD)"
$asset_total_noroof_ppp	"Value of non-land assets excluding roof (USD)"
$asset_land_owned_total	"Land owned (acres)"
$cons_med_total_ppp_m	"Medical expenditure past month (USD)"
$cons_total_ppp	"Total expenditure (USD)"
$ent_total_cost_ppp	"Total expenses, monthly (USD)"
$ent_total_profit_ppp	"Total profit, monthly (USD)"
$b_age	"Age (respondent)"
$b_children	"Number of children"
$b_hhsize	"Household size"
$b_edu	"Years of education completed (respondent)"
$hh_children	"Number of children <=18 or younger in Household"
$hh_workactivities	"Income Generating Activities per HH Adult"
$hh_propsalaried	"Proportion of Adults involved in Wage Labor"
$hh_propcasual	"Proportion of Adults involved in Casual Labor"

Exhibit 2. Important Variable Comparison (slim & wide)

RF (slim)	Boosting (slim)	Boosting (wide)	Lasso (slim)	Boruta (slim)	Lasso (wide)
hh_totalmembers0	hh_totalmembers0	hh_totalmembers0	fs_hhfoodindexnew0_surrogate	fs_hhfoodindexnew0	fs_hhfoodindexnew0
hh_propsalaried0_surrogate	asset_total_ppp1	cons_nondurable_ppp0	b_age_surrogate	fs_hhfoodindexnew1	fs_hhfoodindexnew1
hh_propcasual0_surrogate	med_hh_healthindex0	cons_allfood_ppp_m0	fs_hhfoodindexnew1_surrogate	med_hh_healthindex0	med_hh_healthindex1
hh_workactivities0_surrogate	cons_total_ppp1	nondurable_investment0	fs_childfoodindexnew1_surrogate	med_hh_healthindex1	ed_index0
hh_totalmembers0_surrogate	ed_index0	asset_total_ppp0	ed_index1_surrogate	ed_index0	ed_index1
med_hh_healthindex0_surrogate	ent_total_profit_ppp1	asset_total_ppp1	asset_lntotal_ppp1	ed_index1	psy_index_z1
hh_workactivities0	fs_hhfoodindexnew1	asset_phone_ppp1	hh_totalmembers1	psy_index_z0	ih_overall_index_z0
fs_hhfoodindexnew0_surrogate	med_hh_healthindex1	durable_investment1	med_child_healthindex1_surrogate	psy_index_z1	ih_overall_index_z1
hh_children0	ent_total_cost_ppp1	asset_furniture_ppp0	asset_total_ppp1	ih_overall_index_z0	fs_childfoodindexnew0
asset_total_ppp1	asset_total_noroof_ppp1	cons_total_ppp1	hh_totalmembers0_surrogate	fs_childfoodindexnew0	fs_childfoodindexnew1
				fs_childfoodindexnew1	med_child_healthindex0
				med_child_healthindex0	med_child_healthindex1
				med_child_healthindex1	hh_totalmembers1
				hh_totalmembers0	asset_total_ppp1
				hh_totalmembers1	asset_lntotal_ppp1
				asset_total_ppp1	asset_total_noroof_ppp1
				ent_total_rev_ppp1	asset_land_owned_total0
				asset_lntotal_ppp1	asset_land_owned_total1
				ent_lntotal_rev_ppp1	cons_med_total_ppp_m0
				asset_total_noroof_ppp1	cons_med_total_ppp_m1
				asset_land_owned_total0	cons_total_ppp1
				asset_land_owned_total1	cons_lnmed_total_ppp_m0
				asset_lntotal_noroof_ppp1	cons_lnmed_total_ppp_m1
				cons_med_total_ppp_m0	ent_total_cost_ppp1
				cons_med_total_ppp_m1	ent_total_profit_ppp1
				cons_total_ppp1	ent_lntotal_profit_ppp1

Exhibit 3. Important Variable Comparison (discover)

RF (discover)	Boosting (discover)	Boruta (discover)
b_f_scheier3	b_f_cohen_2	endline_timing
b_f_cesd12	b_m_cohen_2	b_age
b_f_rosenberg4_surrogate	b_m_rosenberg3	b_children
b_f_cesd8	asset_phone_ppp1	b_hhsize
b_f_cesd6	durable_investment1	b_edu
b_f_scheier4	asset_total_ppp1	hh_children1
asset_total_ppp1	b_m_scheier4	hh_totalmembers1
b_f_cesd13	asset_niceroof1	asset_total_ppp1
b_f_cesd2	amount_received_mpesa	cons_nondurable_ppp1
asset_lntotal_ppp1	cons_total_ppp1	ent_total_rev_ppp1
		asset_lntotal_ppp1
		cons_lnondurable_ppp1
		ent_lntotal_rev_ppp1
		asset_total_noroof_ppp1
		asset_livestock_ppp1
		asset_cows_ppp1
		asset_smalllivestock_ppp1
		asset_birds_ppp1
		asset_durable_ppp1
		asset_furniture_ppp1
		asset_ag_ppp1
		asset_radiotv_ppp1
		asset_trans_ppp1
		asset_appliance_ppp1
		asset_phone_ppp1