Reasons for Giving: Urgency and Stability of Matches in Child Sponsorship Programs

Jun Ho Choi

June 12, 2019

Abstract

Using the data collected from Compassion International, this study seeks to answer the question of whether matches in child sponsorship programs are based solely on how urgently the potential sponsorship recipients need help or also influenced by how risky the stability of a match can be. While the information about the potential sponsors are unknown to the researchers, the revealed preference in matches seems to corroborate the claim that the potential sponsors also care about their own satisfaction in a child sponsorship match and avoid to be matched with children who signal that the matches with them could potentially be unstable. This result has been garnered using econometric methods suchas parametric and nonparametric survival regressions and regression discontinuity designs. In addition, the paper seeks to enhance the prediction on the said matches by competing an ensemble of various machine learning methods. In aggregate, both objectives work towards potentially improving upon the current child sponsorship program environments.

1 Introduction

Despite the apparent size of the market for charitable giving and number of people engaging in child-to-sponsor match programs (hereforth CSP), the literature on the said programs is almost barren (List 2011; Compassion 2018). While there are existing studies that focus on how CSP can positively impact the potential sponsorship recipients (hereforth PR), there is little emphasis on why the potential sponsors (hereforth PS) may engage in the CSP through child sponsorship organizations (hereforth CSO) in the first place (Wydick et al. 2013). Equally important, yet also unanswered, question is whether there are attributes from the PR that either make the PS want to or refrain from engaging in matches.

The above-mentioned issues are heavily intertwined with the market design problem of structuring the CSP environment, as the motivation behind trying to understand the factors that may promote or hinder recipient-sponsor matches in a CSP or to make a better prediction of the said matches is essentially to help create a better system in which more recipients can benefit from CSP. From an economist's point of view, this objective equivalent to the goal of maximizing welfare in the *economic system* of CSP.[1]

This paper, therefore, attempts to answer the questions of what are the key covariates in determining match statuses of the PR and what methods can be used to effectively produce future predictions. Specifically, regarding the former question, it seeks to answer whether urgency of needing sponsorship is more important in creating a match than signalling the stability of extending sponsorship, or vice versa. Another way to phrase this, is to say whether the need for sponsorship by PR (and the good will by PS) is more important, or whether the PS also have something to gain from the said matches and seek a more stable relationship with the PR if matched. This idea is based on Andreoni (1989)'s economic theory of warm-glow, in that charitable actions are not driven solely by good will or potential monetary gains but also non-monetary gains, dubbed "warm glow." This theory has garnered support from experimental studies (Andreoni and Miller 2002; Harbaugh et al. 2007).

In doing so, this paper will rely on econometric methods of survival analysis and machine learning methods. As Roth (2002) has argued, in a market design problem, "experimental and computational economics are natural complements to game theory" (p. 1342). And while this paper is not entirely a paper on designing more efficient platforms, it does aim at providing a partial answer to how improvements to the current CSP environment can be made. The said methods will be used in analyzing the dataset gathered from Compassion International (hereforth Compassion), which has information about the PR that the PS observe and base their decisions

[1]This interpretation of CSP as a market economy has been inspired by papers such as Becker and Lewis (1973) and Hansen and Hansen (2006), which focus not on CSP, but on fertility and child adoption, respectively. For a more general discussion on the market for charity, reference can be made to List (2011).

upon. The dataset is a compilation of match statuses and PR information, having a 14-day worth of data (i.e. one initial data gathering and fourteen daily updates).

In essence, this study seeks to present assessments on three different types of "competitions." First is that between parametric and nonparametric regressions in survival analyses settings. This will be done in the attempt to evaluate the explanatory power of different covariates in determining the hazard rate in matching (i.e. how likely it is for an unmatched PR to be matched in the next time interval). Second is the competition among different machine learning methods in improving prediction accuracy and recall on the matches. Doing so will help the CSO identify which are the more likely to get matched and direct resources at promoting those with less likelihood of yield a match. Finally, there is the competition between the convention norm and economic theory behind the motives of charitable actions. Those adhering to former would argue that there is either pecuniary motive or pure good will behind charitable actions; on the other hand, economic theory would elaborate that there may be non-pecuniary personal gains from engaging in charitable actions (Andreoni 1989).

2 Data

2.1 Data Collection and Descriptive Statistics

For this study, I collected the data at the individual level from the official website of Compassion International.[2] The focus was on the various signals about the PR can be picked up by the PS. These sets of information were then analyzed with the expectation that at least some subset of such information can ultimately impact whether a PS commits to a sponsorship (i.e. provides the match). Of course, it is not to argue that the dataset curated for this study is the entirety of the potential covariates and features affecting the PS's decisions. For instance, there may be information from the images of the PS – such as complexion and subtle facial expressions – that may influence the PS, but are not captured by the dataset. However, I also note that effort was made to capture most of the textual information provided to the PS.

Data collection was conducted daily from April 19, 2019 to May 3, 2019 (initial collection and 14 daily follow-ups) using web scraping through Python's Selenium module.[3] Because this dataset has a set window of observation and match/non-match status is not observable after this window, this study effectively is working with right-censored data. Daily web scraping was made at around 9:00 PM Central Standard Time (CST), for approximately 3 hours each session, with

[2]For the up-to-date list of children who are waiting for sponsorship, I refer to the following web page: `https://www.compassion.com/sponsor_a_child/`

[3]I note that while I certainly wished to gather data for longer window of observation, the official Compassion website reassigned the IDs and web URLs associated with each PR on May 4, 2019, making it unable for me to make updates to the existing dataset.

the exception of collections made in May due to network problems, in which the processes started at 11:00 PM CST. The reason for the rather inefficient hours of data collection was due to the structure of the Compassion International's website; information about each individual is stored in each separate web page, rather than the organization having a public list or table of all the individual and corresponding information.

Descriptive statistics of selected variables are listed in Table 1, where I present a cleaned-up version of the dataset, in which issues such as missing variable and encoding problems have been dealth with. The data presented in Table 1 are separated by whether the PR associated with the data were ever matched within the 14-day window of observation. Judging by the difference in means and the t-statistics associated with them, it is observable that there is heterogeneity between the two groups separated by match status.

2.2 A Closer Looker the Data: the Three Main Signals

Among the variables, those located in Panel D of Table 1 require more attention; these variables are the more-prominently signaled information, in which the official website of Compassion highlights using "badges" or indicators next to the photographs of the PR if applicable. Throughout the paper, I assume that (while not making any adjustments to weight on covariates) these three variables are the first-order signals that the PS assesses. In addition, in the urgency-stability dialogue mentioned in the Introduction section, the said variables provide important points of analyses as they can either be signalling urgency or the potential threat to stability. For the sake of convenience, the three variables – urgency signal during the observation window, AIDS-affected area, and vulnerable to exploitation – will be referred to as `urgency`, `AIDS area`, and `exploitation`, for the sake of convenience. [4]

The duality problem of urgency versus stability is even more problematic for the variable `urgency`, which is a signal appears only after 180 days have passed with the status of being *unmatched*. Therefore, in understanding the effect of this variable on match status (which will follow in the Methodology and Data sections), it must be noted that those with `urgency` signals may have had more chances at getting a match due to being exposed on the list for a longer time. Concurrently, it is also possible that `urgency` can have a negative impact for matching as the PS may think there is something awry with the PR with `urgency` status for not being matched for a very long period.

Another problem is the issue of `AIDS area` perfectly predicting those PR from Africa. As observed from Table 1, the statistics for AIDS-affected area and Africa are exactly the same, meaning that the Compassion website has internally coded all those from African nations as exposed

[4] In addition, the corresponding PR's age, status of living with mother, and status of living with father will also be denoted as `age`, `mother`, and `father` hereforth, when applicable.

Table 1: Descriptive Statistics by Match Status

	Matched (1)	Not matched (2)	t-statistic (3)
Panel A. Demographics			
Age	4.760	4.587	5.421
Female	47.22	44.77	2.438
Education			
Too young for enrollment	34.70	32.84	1.917
Unenrolled despite of age	13.44	15.33	−2.641
Preschool and kindergarten	33.86	39.00	−5.229
Elementary and middle schools	18.01	12.83	6.952
Panel B. Family Information			
Living with mother	92.57	91.24	2.397
Living with father	69.70	63.59	6.342
Number of siblings	1.558	1.578	−0.638
Guardian employment status			
Both guardians unemployed or unknown	10.07	12.72	−4.090
Only one guardian employed	51.05	53.71	−2.599
Both employed	38.88	33.57	5.383
Panel C. Geographic Information			
Christianity as the dominant religion	66.67	63.38	3.427
Regional average monthly income (USD)	85.847	73.636	7.703
Continent			
Africa	44.29	52.90	−8.589
Asia	10.50	3.91	12.413
South America	20.51	16.43	5.207
North America and Caribbean	9.22	0.05	8.533
Central America	15.49	21.98	−8.361
Panel D. Signaling Information			
Urgency (during the observation window)	4.62	0.11	14.115
AIDS-affected area	44.29	52.90	−8.589
Vulnerable to exploitation	35.66	43.63	−8.128
N	4132	5386	-

NOTE: Entries for categorical variables are in percentage terms, and those for continuous variables are averaged by group. Education levels are indicated as those that correspond to the U.S. educational system. The dummy variable "Christianity as the dominant religion" equals to 1 if the corresponding PR's nation has 70% or more population designated as Christians, and is considered as Compassion International is a Christian non-profit organization.

to the threat of AIDS/HIV. I note that this problem is not observable if the dataset has not been carefully studied; in other words, it is unlikely that a PS would think *all* of Africa is exposed to the threat of AIDS. In fact there are African nations with lower HIV/AIDS prevalence than some of the non-African nations; for instance, Burkina Faso, in which 10.79% of the overall sample PR are located, has a lower HIV/AIDS adult prevalence rate than, say, Haiti, which is not indicated as AIDS-affected in the dataset (CIA n.d.). Therefore, while there is no way to disentangle the

two variables (which are effectively one variable), one can exert additional effort in interpreting the results with respect to the PR in African nations.

3 Methodology

This study implements econometric and machine learning analyses, in which both broods of methodology will be working with the right-censored data for child-to-sponsor matches. Despite the differences, both methods are used to understand how a match in the CSP setting is made and to produce results that can eventually be utilized to improve upon the current CSP environment.

The econometric models are aimed at understanding how the various information about the PR that CSO signals to the PS can potentially impact creation of matches. Implementation of such models and interpreting the results from them will allow one to evaluate the aforementioned competing theories on why the PS may engage in CSP. On the other hand, the machine learning methods will be focused at producing accurate predictions and gauging which features are influential in the classification, thus potentially corroborating or refuting the results from econometric models.

3.1 Econometric Methods

Survival Regression Analyses — Using the right-censored data collected from Compassion, survival regressions will be the main strategy that this paper employs in order to understand which variables may affect making of matches (or the lack of one) in a CSP environment. One of the key objects of interest is how long it would take for a match to be made given the covariates. This maps directly to what the survival regressions target at modeling, which is the conditional hazard (i.e. the probability that an observation reaches "death" in the following period given that one is "alive" at the current period and the covariates) (Ciuca and Matei 2010).

This paper explores using both parametric and nonparametric survival regressions, following Cox (1972). For the parametric version, this paper will showcase the use of Weibull regression; the nonparametric case is also known as Cox regression (Rodriguez 2010). The two cases differ only by how the baseline hazard function is estimated; in the parametric case, the baseline hazard function assumes a predefined functional form (e.g. Weibull regression with the baseline hazard function as Weibull hazard function), whereas in the nonparametric (or semiparametric) case it is estimated using the data (Rodriguez 2010).

Despite the seemingly-mild differences, both nonparametric and parametric cases are considered for the sake of robustness in the analysis. Parametric survival regressions, due to assuming a specific baseline hazard function, may produce less realistic results. For instance, the conditional hazard

functions (to be defined below) produced by Weibull regressions will only take on monotonic forms. Nonparametric versions do not have this problem, but because the baseline is not specified, it may have the problem of overfitting. The two methods, therefore, are complimentary of one another this analysis.

In both parametric and nonparametric survival regressions, the model is as follows:

$$\lambda_i(t|X_i) = \lambda_0(t) \exp(X_i'\beta)$$

where $\lambda_0(t)$ is the baseline hazard function at time t, X_i' is the vector of covariates for individual i, and $\lambda_i(t|X_i)$ is the conditional hazard function at time t given the said covariates. Notice that because we are not actually able to observe data points at continuous time, we will actually be estimating conditional hazard in a by-parts manner, so that

$$\lambda_{ij}(X_i) = \lambda_j \exp(X_i'\beta)$$

where j denotes time *interval*, and λ_{ij} and λ_j are the by-parts analogue of the continuous-time conditional hazard and baseline hazard functions.

For the covariates (X_i), this study will primarily consider using the three main signals (i.e. `urgency`, `AIDS area`, and `expropriation`) with the control variables as `age`, `female`, `father`, `mother`, `page` dummies, continent dummies, guardian employment statuses, and educational levels of the PR. The three main signals are of primary interest as these are what the PS can initially observe and what Compassion highlights. Note, however, that `urgency` requires further analysis due to being a "mixed signal," and therefore I will conduct regression discontinuity analysis for this variable (explained further below).

The said control variables are placed to better learn of the effect of the three signals on making CSP matches. However, controls such as guardian employment statuses have the potential to be interpreted as signalling stability of match for the PS, and therefore may need further post-estimation analysis. Additional controls, such as average monthly income of the region and number of siblings, will be noted when used.

Note on Standard Errors and Multiple Hypotheses Testing — As the `page`s are ordered by how long the PR have been waiting, and because the PS observe the PR in a `page`-by-`page` manner, it can be assumed that there may be structural unobservables correlated with one another at the `page` level. Therefore, this study considers the use of clustered standard errors (SE) at the `page` level.

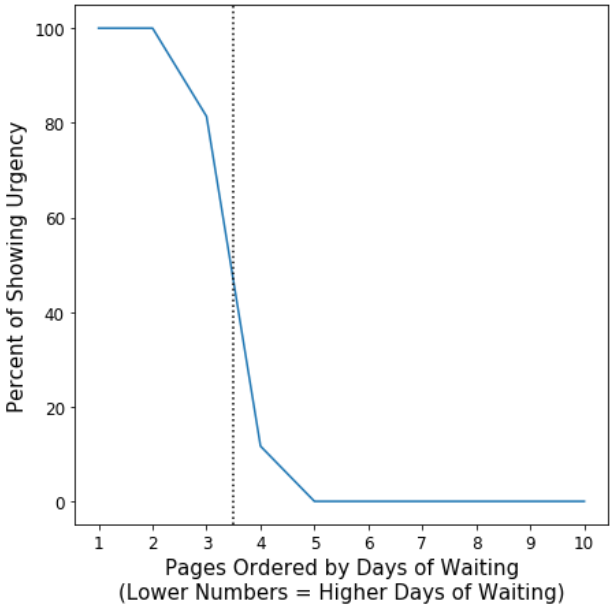
In addition, because the PS are observing multiple signals at the same time and such signals are contemporaneously affecting decision-making by the PS, one must consider testing multiple hypotheses at the same time instead of focusing on individual estimates and their statistical sig-

nificance. One easy way of implementing this is to use the Bonferroni correction, which is to test each individual hypothesis at the significance level of α/K when the target significance is at α and there are K hypotheses to test (Dunn 1961).

Regression Discontinuity Design (RDD) for `Urgency` — As mentioned in the Data section, `urgency` is a "mixed signal" in that it informs the PS about not only how dire the need for help is for a certain PR, but also how long the said PR has been on the CSP list and therefore has had more chances at getting a match. One way of disentangling this duality would be to look at the effect of `urgency` just at the threshold where it turns on – that is, the conditional average treatment effect (CATE) at the threshold.

The problem is, while the said threshold is exactly 180 days after being on the list, the days of waiting are unobservable to the PS when 180 days have not passed. However, the dataset does have a proxy for the days of waiting, which is the page order (on the Compassion website) by days of waiting (i.e. the longer-waiting PR would be placed in the earlier pages).[5] As seen from Figure 1, there is a monotonic increase of percent showing `urgency` between pages 3 and 4; this motivates the use of fuzzy RDD, in which the probability of being "treated by `urgency`" does not jump from 0 to 1 (as in the sharp RDD setting), but there is still a sizable, discontinuous change in the treatment probability at the threshold.

Figure 1: Percent of Showing Urgency by Pages



[5]I note that the method of assuming a parametric distribution for the days of waiting and fitting the known data points using maximum likelihood estimation was also considered. However, because the days of waiting are observed only for the first 3 pages (approximately only 2% out of the total 140 pages), I deemed that this would be highly inaccurate.

In the context of this paper, the said threshold is page 3. I follow Hahn, Todd, Van der Klaauw (2001) and use the instrumental-variables version of fuzzy RDD, which is to estimate the following regression:

$$Y_i = \alpha + \tau \texttt{urgency}_i + \beta(\texttt{page}_i - 3) + \delta Z_i(\texttt{page}_i - 3) + \varepsilon_i$$

where Y_i is the match status (= 1 if matched during the window of observation), $\texttt{urgency}_i$ is instrumented by $Z_i \equiv 1\{\texttt{page} \leq 3\}$, and α, τ, β, and δ are estimated by local linear regression as follows:

$$(\hat{\alpha}, \hat{\tau}, \hat{\beta}, \hat{\delta}) = \underset{\alpha,\tau,\beta,\delta}{\operatorname{argmin}} \sum_{i=1}^{n} \big(Y_i - \alpha - \tau \texttt{urgency}_i - \delta Z_i(\texttt{page}_i - 3)\big)^2 K_h(\texttt{page}_i - 3)$$

in which $K_h(\cdot)$ is the kernel function with bandwidth h. In this paper, we use triangular kernel with optimal bandwidth h selected by the method of Calonico, Cattaneo, and Titiunik (2014). In addition to conducting fuzzy RDD where threshold is identified as `page` 3, similar designs using `page` 2 and `page` 4 as thresholds will be conducted for the sake of sensitivity analyses.

Note that because the fuzzy RDD described above utilizes IV, it must satisfy relevance and exclusion restriction. Relevance follows from there being a threshold for being "treated" with `urgency`. Exclusion restriction is tricky, as `page` itself may be affecting match status; for instance, it may be that the PS have less concern for examining the PR in the latter pages than earlier pages, due to convenience. However, because the problem at hand is estimating the CATE at the threshold, I argue that `page` would not affect the PR decision locally; that is, for instance, the PR on `page` 3 and those on `page` 4 are roughly similar in being chosen with all else equal. This assumption would help one apply exclusion restriction in this setting.

3.2 Machine Learning Methods

Survival Decision Tree and Random Forest — While newer and more sophisticated machine learning (ML) methods are being developed or already in use, I argue in this section that decision tree and random forest methods are suitable for the question of predicting CSP match status using the right-censored data at hand (Bou-Hamad et al. 2011). For this argument, I make exposition that the maximum likelihood estimation (MLE) using hazard function coincides with decision tree and random forest methods using information gain (or entropy) as a splitting criterion; the below process closely follows Kuhn (2014).

As in the survival regression methods, it is most likely that one is unable to observe continuous-time status of survival from many datasets. Therefore, the log-likelihood of MLE of optimizing for hazard functions associated with the survival of an observation i at time interval T will take the

following form:

$$L_i(T, \theta_i|X_i) = \log\Big[\Big(\prod_{t=1}^{T-1} \lambda(t|X_i)\Big)\Big(\lambda(T|X_i)^{\theta_i}\{1-\lambda(T|X_i)\}^{(1-\theta_i)}\Big)\Big]$$

in which $t \in \{1, 2, \ldots, T\}$ is a time interval, X_i are the covariates, λ the population conditional hazard function, θ_i the status of i being alive (= 1) or dead (= 0), and finally L_i the conditional log-likelihood function. Reorganization of the terms (using log-transformations) yield

$$L_i(T, \theta_i|X_i, \hat{\lambda}) = \Big[(1-\theta_i)\log\{1-\hat{\lambda}(T|X_i)\} + \theta_i \log\hat{\lambda}(t|X_i) + \sum_{t=1}^{T-1}\log\hat{\lambda}(t|X_i)\Big]$$

in which λ has been swapped with its sample analogue, $\hat{\lambda}$ for estimation. Notice that the goal of MLE over all the samples is to optimize (maximize) the average log-likelihood. At the same time, because for each time interval t, observation i is either alive or dead, we can denote t-specific status as θ_{it}. Denoting the average log-likelihood over the entire sample as A, we can write

$$A(T, \theta|X, \hat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\Big[(1-\theta_{it})\log\{1-\hat{\lambda}(t|X_i)\} + \theta_{it}\log\hat{\lambda}(t|X_i)\Big] \tag{A}$$

where now, end period (due to right-censoring or reaching death) is different for each observation i and therefore is denoted with subscript i (i.e. T_i). Notice, however, that this is essentially a classification problem over all the θ_{it}'s, in which one tries to predict the binary status for each i and t combination. The said classification problem is implementable in a decision tree setting because the information gain splitting criterion has the exact same form as that in the expression (A) above. One can also expand application to random forest by using the same splitting criterion.

As the said methods are applicable to the survival analysis setting, I will refer to them as survival decision tree (hereforth SDT) and survival random forest (hereforth SRF) respectively; both methods will be used in a multi-class classification problem in which the object of prediction is the status of observations at different time intervals (i.e. θ_{it} above). However, due to the short time-dimension in the dataset, I consider the very simple problem of classifying whether a PR is alive (i.e. does not match) during the entire window of observation or fails (i.e. does match) at some point.

Note on Features Used and Criterion for Optimization — The sets of features used will be described as either "whole" or "feature-selected," and will subsequently referred to as such. The whole version uses as many features as possible in the dataset, while the feature-selected version uses a smaller subset of features that are used in the survival regression analyses above. The former is to gain higher predictive accuracy and recall for the matches (to be discussed below), and the latter is to implement feature importance analysis to observe whether the features important for

predictions using SRF and SDT align with those shown as statistically significant in the regression analyses. The lists of features used in the said versions are summarized in Table 2.

Table 2: Features Used in the Machine Learning Methods

Version Designation	List of Features
Feature-selected	`Urgency`, `age`, `female`, `AIDS area`, `exploitation`, `mother`, `father`, guardian employment status dummies, education status dummies, continent dummies
Whole	All features used in the feature-selected version, and additionally `page` dummies, country dummies, and language dummies

This study also implements hyperparameter tuning (HPT) and k-folds cross-validation for increasing predictive accuracy and recall for the matches. Specifically, I aim at maximizing recall, which is defined as the ratio of true positives to relative elements (i.e. the sum of true positives and false negatives) (Powers 2011). That is, the goal is to find as many observations predicted as matched (unmatched) out of those that are actually matched (unmatched). While overall accuracy, too, is an important metric (and therefore reported in the Results section), I argue that by optimizing for recall we can better direct resources of the CSO in their attempt to match those that are unlikely to be matched. The PR with attractive features will not have problems in yielding a match; on the other hand, the PR without such may require further spending or advertisements to successfully get one. By optimizing for recall, the CSO would quickly be able to identify those that are unlikely to get matches and embark on additional pushes for the said PR.

Additional ML Methods for Improving Prediction — While I have argued that decision trees or random forests will be most optimal for the prediction, this paper will also use additional methods and compare the prediction results using metrics such as accuracy, recall on the matches, plotting of the receiver operating curve (ROC), and area under the ROC curve (AUC). Table 3 lists and provides brief details for the said additional methods used; these methods were chosen to attain some balance between linear and non-linear models. When applicable, the hyperparameter tuning via grid search or random search method and k-folds crossvalidation will be conducted as in the case for survival decision trees and survival random forests. Finally, the features used in conjunction with these methods are the same as those designated for survival decision tree and survival random forest.

Table 3: Additional Machine Learning Methods Used in the Study

Classifier	Abbrev.	Lin.	HPT Strategy
Linear Discriminant Analysis	LinDA	Y	Shrinkage parameter
k-Nearest Neighbors	KNN	N	k, distance metric
Multilayer Perceptron	MLP	N	Hidden layer size, $l2$, activation fn.
Support Vector Machine	SVM	Y	Kernel, penalty parameter (C)

NOTE: For linear discriminant analysis, this study uses the least-squares solver. "Lin." refers to whether a method is a linear one or not. HPT and CV refer to hyperparameter tuning and k-folds cross-validation, respectively. "fn." refers to function.

4 Results

4.1 Econometric Methods Results

Survival Regression Results and the Bonferroni Correction — Presented in Table 4 are the hazard rate estimates (with `page`-clustered SEs) from both Cox and Weibull regressions. Despite the fact that one assumes a functional form for the baseline hazard rate function and the other does not, the results are well aligned. The three main signals – `urgency`, `AIDS area`, and `exploitation` – all yield statistically significant results at the 1% significance level. Note that I will subsequently carry out the Bonferroni correction to see whether their statistical significance holds when testing for multiple hypotheses.

To elaborate further on the **three main signals**, that the hazard rate estimates are less one mean that having the said signals actually improves the survival rate – that is, *decreases* the probability of matching. With respect to `AIDS area` and `exploitation`, this is in alignment with Andreoni (1989) in that the PS are seeking PR that are going to produce stable matches for an extended time, so that they themselves can receive utility from the match. However, due to the duality of `urgency` that was elaborated in the Data section, it is difficult to say for certain whether those with `urgency` signals are less likely to get matched. Due to the regressions being linear in `urgency`, it could be that it does not fully capture the non-linear underlying structure it has. Another alternative explanation is that, due to the higher correlation between `urgency` and `page`s, the explantory power is absorbed by the latter (which, in the form of dummy variables, are included in the regression as well). This unresolved issue with `urgency` will be explored more deeply with fuzzy RDD.

Other points of interest, which show statistical significance across the regressions, are as follows: **female** PR are more likely to quickly yield matches in comparison to male PR; in addition, that those with **only a single guardian being employed** are less likely to be matched over both guardians employed *and* all guardians being unemployed or unknown. Regarding the preference for female PR, one could infer from similar tendencies in child adoption, in which preference for

Table 4: Survival Regression Results

	Cox		Weibull	
	(1)	(2)	(3)	(4)
`Urgency`	0.4450***	0.4460***	0.3901***	0.3959***
	(0.0650)	(0.0621)	(0.0596)	(0.0579)
`AIDS area` (Africa)	0.5344***	0.6152***	0.5212***	0.6074***
	(0.0824)	(0.1012)	(0.0899)	(0.1100)
`Exploitation`	0.6124***	0.6609***	0.5914***	0.6477***
	(0.0981)	(0.1033)	(0.1036)	(0.1105)
`Age`	1.0355	1.0368	1.0443	1.0454
	(0.0360)	(0.0369)	(0.0394)	(0.0406)
Female	1.2039***	1.2090***	1.2183***	1.2255***
	(0.0362)	(0.0370)	(0.0391)	(0.0403)
Living with mother	1.0268	1.0112	1.0322	1.0147
	(0.0598)	(0.0691)	(0.0632)	(0.0728)
Living with father	1.0111	1.0160	1.0104	1.0162
	(0.0387)	(0.0416)	(0.0410)	(0.0440)
One guardian employed	0.8433**	0.8599**	0.8274**	0.8478**
	(0.0662)	(0.0646)	(0.0708)	(0.0687)
Both guardians employed	1.0954	1.1244	1.0837	1.1201
	(0.1039)	(0.0984)	(0.1105)	(0.1041)
Unenrolled	0.9452	0.9437	0.9389	0.9390
	(0.0794)	(0.0783)	(0.0844)	(0.0834)
Preschool & Kindergarten	0.9672	0.9619	0.9684	0.9601
	(0.0730)	(0.0694)	(0.0786)	(0.0739)
Elementary or Middle	1.0013	1.0004	1.0008	0.9994
	(0.1011)	(0.0984)	(0.1083)	(0.1054)
Regional monthly income		1.0009*		1.0009*
		(0.0005)		(0.0005)
N	9518	9518	9518	9518
`Page` dummies	Yes	Yes	Yes	Yes
Continent dummies	Yes	Yes	Yes	Yes
Additional controls	No	Yes	No	Yes

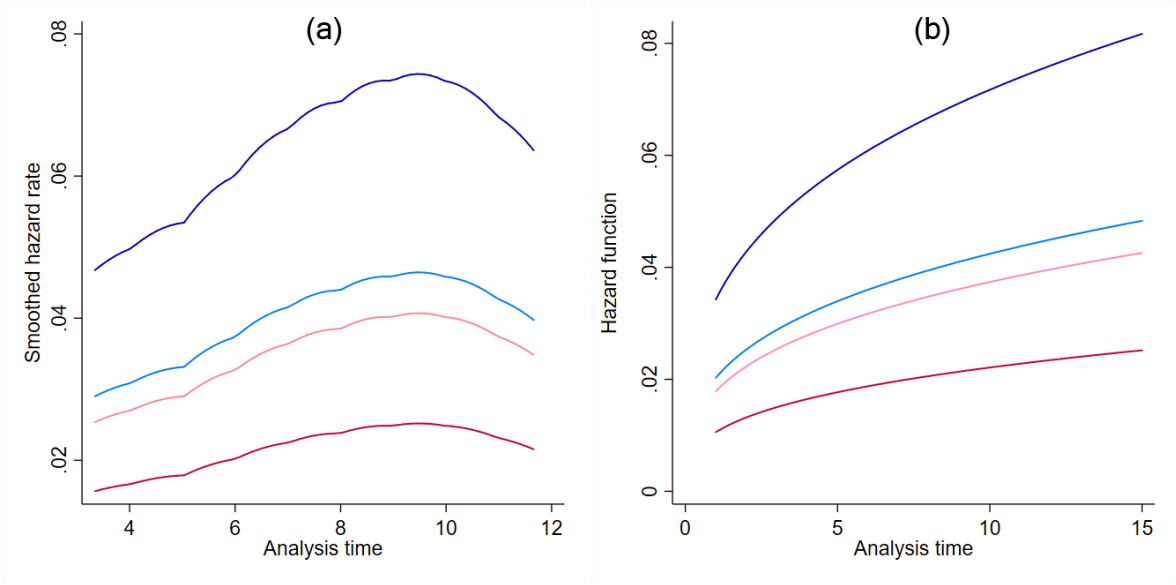
NOTE: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$. Heteroskedasticity-robust SE in parentheses. Hazard rate estimates reported for all columns, in which the dependent variable is the conditional hazard rate. Columns (1) and (2) report results from Cox regression; (3) and (4) report those from Weibull regression. Values less than 1 indicate reducing the hazard rate (i.e. correlated with longer period of not being matched). Baseline for education is "too young to be in school"; that for guardian employment is "both unemployed or unknown." `Page` dummies have the baseline of `page` 1. Continent dummies include Asia, South America, Central America, with the baseline of North America/Caribbean. Additional controls include regional monthly income (scaled to USD), percent of Christians in a country, number of siblings in a family, and living with grandparents. Constant omitted for Weibull regressions.

female children are shown as well. In Gravois (2004), the author posits the possibility that the majority of adoptive parents or interested parties are female, and that there is a tendency for adoptive parents to choose the child of the same biological sex than to do the otherwise. Baccara et al. (2014) also show that in child-to-adoptive parents matching, there is preference for certain ethnicities and genders (African-American and females). Perhaps a similar phenomenon is occurring

here, in which there are more female PS than male ones, and they are more willing to adopt female PR. However, because I have not been able to acquire data on PS demographics, this idea remains to be a hypothesis.

Similar, inconclusive yet exploratory, hypothesis may be made regarding the statistical significance on conditional hazard of only one guardian employed being less than 1. From the stability dialogue, one may compare this estimate with that of both guardians being employed and elaborate that this can mean that the PS look for safer matches in which the sponsorship is simply given as a slight increment to the PR's household budget, and the main source of income is stably coming from both parents or guardians working. On the other hand, one may adhere to the urgency dialogue, compare the exact same estimate with the baseline of no guardians working (or unknown), and arrive at the idea that the PS are willing to help the most needy. It is most likely that both urgency and stability are at work; due to this, there could be some multiple equilibria in which, rather ironically, those with *some* source of income (from one parent working) are less likely to quickly get matched in comparison to those with either *zero* or *more* source of income.

Figure 2: Hazard Rate Comparison of Cox and Weibull Regressions



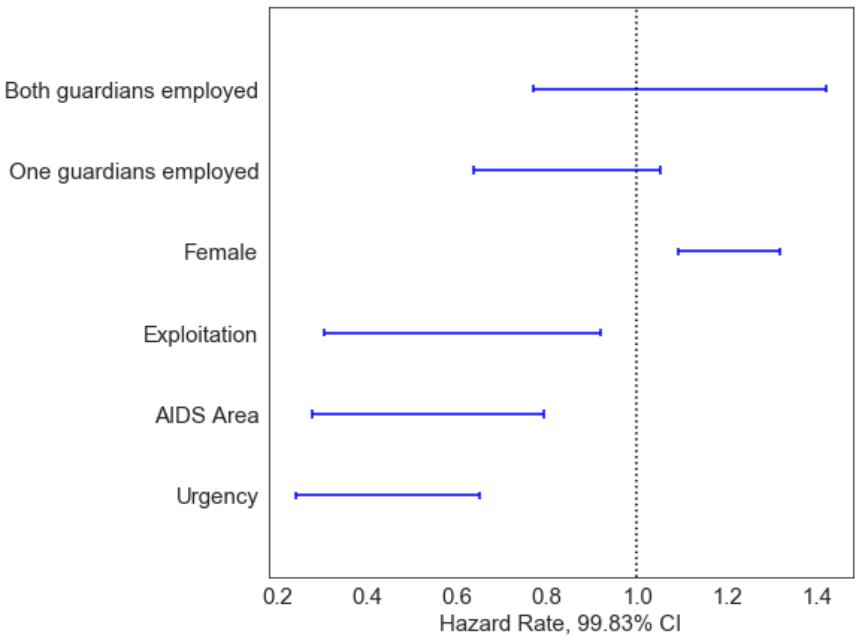
NOTE: Panels (a) and (b) each refer to smoothed hazard rate and hazard calculated from Cox and Weibull regressions. From top to bottom in both panels, the lines correspond to neither under `expropriation` nor in `AIDS area`, under `expropriation` but not in `AIDS area`, not under `expropriation` but in `AIDS area`, and finally under both.

Before presenting the robustness under multiple testing, I compare the smoothed hazard rate (of Cox regression) and the fitted hazard function (of Weibull regression) in Figure 2 in which panels (a) and (b) refer to results from Cox and Weibull regressions, respectively. As expectedly, the hazard function of Weibull regression is monotonically increasing in time. In addition, in both cases, when covariates associated with a lesser likelihood of match are stacked, the hazard function

or smoothed hazard rate values are globally lower (within the context of this analysis). However, judging from the shape of the hazard functions, it seems that the smoothed hazard rate from Cox regression is more realistic. This is due to the observation that in the dataset, there are those with extremely long days of waiting. If hazard functions were monotonic, the cumulative hazard would be growing very fast and it would be extremely unlikely to observe those with long days of waiting. Therefore, I conclude that nonparametric Cox regression is more suitable for analysis.

All of this discussion may be at risk if it is shown that there is no statistical significance for the multiple testing of hypotheses; this is because, as elaborated in the Methodology section, the PS observe multiple signals at the same time. I employ Bonferroni correction for testing 6 hypotheses at the same time, which is going to test, contemporaneously, the hypotheses that the estimated hazard rates for the three main signals, female binary variable, and the guardian employment statuses variable are different from 1. These covariates have chosen as either they were statistically significant at the 5% level or are associated with closely with the said significant variables (i.e. both guardians employed being related to only one guardian being employed). The benchmark statistical significance is 1%, and therefore this would be roughly similar to testing each hypothesis at approximately 0.167% significance level (and constructing 99.83% confidence interval for each).[6]

Figure 3: Multiple Testing with Bonferroni-Corrected Confidence Intervals



NOTE: Results used to create the confidence intervals are shown in Cox and Weibull regressions. From top to bottom in both panels, the lines correspond to neither under `expropriation` nor in `AIDS area`, under `expropriation` but not in `AIDS area`, not under `expropriation` but in `AIDS area`, and finally under both.

[6]Of course, a "less harsh" benchmark of 5% statistical significance could be used, but I wanted to push the test so that only relevant covariates for analysis can be focused upon.

The resulting 99.83% confidence intervals are plotted in Figure 3, in which the results have been created from Cox regression without added controls (i.e. corresponding to column (1) in Table 4). It can be seen that the guardian employment status is not statistically significant enough to influence match statuses; the surviving covariates are the three signals along with female binary variable. The interpretations for the said four variables remain the same.

Fuzzy RDD Results for `Urgency` — As elaborated in the Methodology section, fuzzy RDD around `page` 3 (and around `page` 2 and `page` 4 for sensitivity) is conducted in order to address concerns regarding the duality of `urgency`. The results are produced in Table 5.

Table 5: Results for Fuzzy Regression Discontinuity Design (Using IV)

	Threshold `page`		
	`page` 3	`page` 2	`page` 4
Panel A. Structural Estimates			
`Urgency`	0.1067^{***}	-0.0194	-0.2142
	(0.0278)	(0.0469)	(0.3236)
Panel B. First-stage Estimates			
`Page`	-0.8179^{***}	-0.3412^{***}	-0.2190
	(0.0331)	(0.0570)	(0.1477)
Total N	9518	9518	9518
Left N	109	45	168
Right N	950	666	261
Optimal bandwidth	13.093	9.150	3.445

NOTE: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$. Heteroskedasticity-robust SE in parentheses. Out of the three threshold `page`s, using `page` 3 produces the main result; others are used for sensitivity analysis. Dependent variable is binary match status during the window of observation (i.e. = 1 if matched in the window). Left and right N refer to the numbers of observations with `page` less than or equal to the threshold and greater than the threshold used for fitting the local linear regression. Triangle kernel function was used.

I begin by examining the first-stage estimates (Panel B in Table 5), which will gauge how strong the relationship between `page` and `urgency`. Notice that while the values are all negative – which indicates that the higher the `page` number, the less likely it is to observe `urgency` –, the statistical significance as well as the magnitude of coefficient are the greatest for the threshold of `page` 3. This is expected, as verifiable also from Figure 1.

Next, I examine the CATE of `urgency` on match status at the designated thresholds (Panel A in Table 5). Noting that positive values mean that `urgency` signal is positively correlated with match status at given thresholds, it is confirmable that the only positive and statistically significant result is yielded when at the threshold of `page` 3. Therefore, despite the worry for duality, there is significant evidence that `urgency` signal is positively affecting matches at the threshold where the said signal turns on.

However, I emphasize that this is certainly not a global estimate of `urgency`'s impact on match-

ing. The above analysis provides ground for clearing the said signal's duality problem at the threshold. Nonetheless, one may still hypothesize that if `urgency` signal is displayed for an elongated time, this leaves negative impression for the PS and therefore matches negatively affected. The said hypothesis, which pinpoints the lingering concern, may explain why the coefficients on `urgency` was less than 1 (with statistical significance) for the survival regression analyses where it addresses the *global* average effect of the covariates.[7]

4.2 Machine Learning Methods Results

Survival Decision Tree and Survival Random Forest — In this part, I present the feature importance comparison of SDT and SRF, both their whole and feature-selected versions. This is to confirm that the covariates in the regression analyses that had statistical significance are also influential in making prediction in the survival machine learning methods. Regarding the prediction accuracy and recall on the test set, I present the said results together with other methods in the below section.

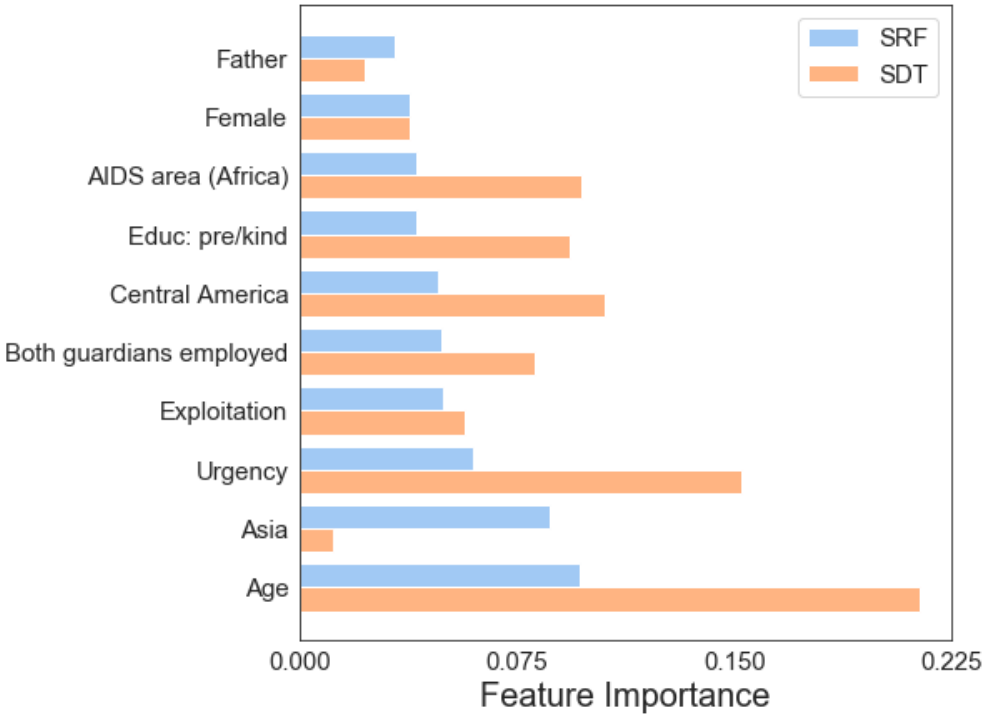
Presented in Figure 4 are the top ten features in the featured-selected version of SRF with the highest feature importance, and their feature importance scores. Plotted together are the corresponding feature importance scores for the same features in the feature-selected version of SDT. It is observable that variables such as `age` and the three main signals are shown to have higher feature importance scores in both methods. The scores for SRF are more nuanced, most likely due to SRF being an ensemble of different SDT.

It is rather unexpected, however, that features such as languages (e.g. Oromiffa and Swahili) and individual countries (e.g. Ethiopia and Philippines) are more important than the three main signals in the whole versions of both methods; this is verifiable in Table 6 in which the top five features with higher feature importance scores are listed in a descending manner, for each method. One hypothesis for this result is that there is highly correlation between the features, such as individual countries and languages, and the key covariates designated in the regression analyses, and that the former features take away the predictive power from the latter as they have more specificity. Another, while less likely, potential reason is that there can be many unobservable or latent features that influence the decisions of PS regarding creation of matches, and by including more observable features into estimation the unobservables being picked up.

Comparison with Other Machine Learning Methods — As aforementioned in the Methodology section, comparisons across different methods will be made using metrics such as predictive accuracy and recall on the CSP matches. Table 7 showcases the results, as well as the brief details on how the hyperparameters were tuned in the optimal case.

[7] Once again, an estimate of less than 1 (more than 1) in the survival regressions maps to a negative (positive) estimate in OLS or logistic regressions.

Figure 4: Feature Importance Comparison: SRF and SDT



NOTE: Only 10 features with highest feature importance from feature-selected survival random forest are plotted. Both survival models are in their feature-selected versions. SDT and SRF refer to survival decision tree and survival random forest, respectively.

Table 6: Top Five Features with Higher Feature Importance, SDT and SRF

Method	Version	Top Five in Feature Importance, Descending Order
SDT	Whole	Oromiffa, Uganda, Swahili, `page` 4, `exploitation`
	FS	`age`, `urgency`, Central America, `AIDS area`, Education: P/K
SRF	Whole	Ethiopia, Philippines, Amharic, French, `page` 2
	FS	`age`, Asia, `urgency`, `exploitation`, both guardians employed

NOTE: The version designation of either "whole" or "feature selected" (abbreviated to FS above) is based on the features used in estimation; the lists for the said features are given in Table 2. "Education: P/K" refers to the binary variable of the PR attending either preschool or kindergarten.

A key point of observation from Table 7 is that the predictive accuracy as well as recall for the matches increase in the whole version of the methods. There are exceptions for SRF and SDT in which the former has a slightly-reduced recall on the matches where as the latter's accuracy dramatically decreases. This suggests that in the whole version case, SDT is not performing well and rather overfitting using the training set.

In the whole version case, MLP shows outstanding performance in all of the metrics I am using for comparison, implying that neural network models are suitable when there is an abundance of data. On the other hand, SRF seems to work well when there are fewer features used; this is most

Table 7: Comparison of Accuracy, Recall for the Matches, and AUC

Method	Accuracy	Recall (= 1)	AUC	Brief HPT Details
Panel A: Whole version				
SDT	0.5823	0.4934	0.6231	Minimum samples per leaf: 8
SRF	0.6811	0.4501	0.7180	Minimum samples per leaf: 5
LinDA	0.6420	0.6751	0.6868	Shrinkage: 0.95
KNN	0.6702	0.5744	0.7087	k: 3, metric: Euclidean
MLP	**0.7361**	**0.6846**	**0.8103**	Hidden layer size: 88, $l2$: 0.6
SVM	0.6966	0.6582	0.7731	Kernel: linear, C: 3.5939
Panel B: Feature-selected version				
SDT	0.6550	0.4633	0.6708	Minimum samples per leaf: 8
SRF	**0.6588**	0.4595	**0.6915**	Minimum samples per leaf: 11
LinDA	0.5840	0.3578	0.5973	Shrinkage: 0.90
KNN	0.6269	0.5113	0.6582	k: 9, metric: Minkowski
MLP	0.6294	**0.5471**	0.6842	Hidden layer size: 23, $l2$: 0.1
SVM	0.5118	0.4218	0.4865	Kernel: sigmoid, C: 1.0081

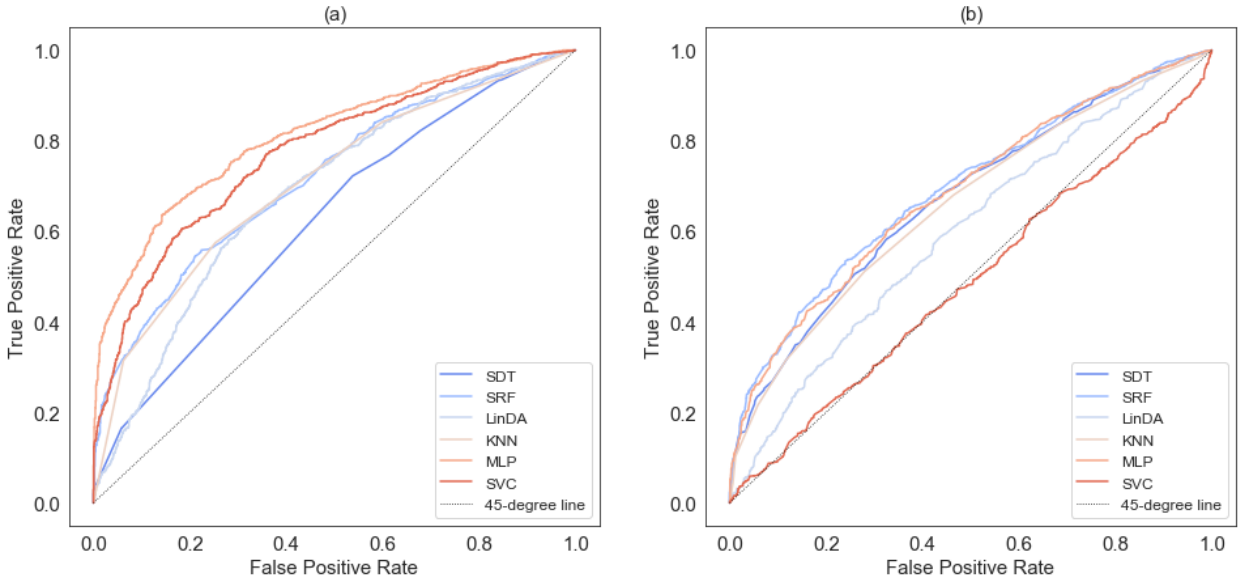
NOTE: In bold are the maximum values for each column (by panel). The version designation of either "whole" or "feature selected" (abbreviated to FS above) is based on the features used in estimation; the lists for the said features are given in Table 2. Accuracy and Recall (= 1) calculated based on the prediction on the test set. "Recall (= 1)" refers to the ratio of true positives to (true positives + false negatives) for the CSP matches. "AUC" refers to the area under the ROC curve, calculated using the test set.

likely due to us using a right-censored data set, as argued before. Still, MLP does well in correctly predicting matches out of those that were actually matched (i.e. higher recall for the matches) even for the feature-selected version. Therefore, the results tell us that MLP is a reliable method in general, while SRF is suggested only when working with limited information. On the other hand, SVM is absolutely not to be used when there is a small number of features, despite its high performance with high-dimensional data. Figure 5, showing the ROC curve plots, also corroborates the above findings.

5 Conclusion

This study has used the data gathered from Compassion Internal to conduct econometric and machine learning analyses for identifying the important covariates in creating matches as well as for producing predictions with high accuracy and recall for matches. From the econometric analyses – which mainly employed survival regressions –, the study has revealed that, on average, three main signals of the PR – urgency, living in AIDS-affected area, and vulnerable to exploitation – are negatively related to quickly yielding matches in a CSP environment. At the same time, at the threshold of the `urgency` signal, fuzzy regression discontinuity design has revealed that it may have a positive impact on creating matches, suggesting that perhaps there is a nonlinear relationship between `urgency` and creation of matches. These results give weight to the warm-glow theory of giving, which hypothesizes that donors to charities (or sponsors, in this context) have personal,

Figure 5: Receiver Operatoring Characteristic (ROC) Curves for Various Methods



NOTE: Panels (a) and (b) each refer to the "whole" and "feature-selected" versions, respectively. Table 2 lists the features used in each version. SDT, SRF, LinDA, KNN, MLP, and SVC each refer to survival decision tree, survival random forest, linear discriminant analysis, k-nearest neighbors, multilayer perceptron, and support vector machine classifier, respectively.

non-pecuniary gains from engaging in charitable activities (Andreoni 1989).

Through the machine learning analyses, this study has arrived at the conclusion of recommending multilayer perceptron or other neural network methods in predicting match statuses. This is given the assumption that Compassion International or similar CSO have higher-dimensional data to work with. In case they do not, using survival random forest can also be recommended for producing predictions as the dataset for match statuses are structured to be right-censored. Mathematical exposition for why survival analysis maps to random forest or decision tree with information gain splitting criterion has been given in the Methodology section.

I end by pointing to some of the caveats of the study and future directions that researchers can take on. Firstly, because the dataset this paper utilized does not have direct information about the potential sponsors, one could only infer their general preferences from the matches. Furthermore, the time dimension is very short for the dataset to be utilized in multi-class classification exercises. In the future, it may be wise to directly work with a CSO and acquire higher-quality data, under suitable circumstances. Secondly, the study could be benefited from developing a simple, theoretical model of how bilateral matches or trade agreements are created based on the potential buyers observing signals from potential sellers, in which the signals are mediated by an intermediary platform. This theoretical exercise is one of the future directions that I am embarking on, where motivations are coming from papers such as Myerson and Satterthwaite (1983).

Reference

Andreoni, James. 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy* 97 (6): 1447–1458.

Andreoni, James, and John Miller. 2002. "Giving According to GARP: An Experimental Test to the Consistency of Preferences for Altruism." *Econometrica* 70 (2): 737–753, March.

Baccara, Mariagiovanna, Allan Collard-Wexler, Leonardo Felli, and Leeat Yariv. 2014. "Child-Adoption Matching: Preferences for Gender and Race." *American Economic Journal: Applied Economics* 6 (3): 133-158.

Becker, Gary S., and H. Gregg Lewis. 1973. "On the Interaction between the Quantity and Quality of Children." *Journal of Political Economy* 81 (2): S279-S288, March-April.

Bou-Hamad, Imad, Denis Larocque, and Hatem Ben-Ameur. 2011. "A Review of Survival Trees." *Statistics Surveys* 5: 44-71.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica*, 82 (6): 2295-2326, November.

Ciuca, Vasilica, and Monica Matei. 2010. "Survival Analysis for the Unemployment Duration." *Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation* 1: 354-359.

Compassion International. 2018. "Annual Report 2017-2018." `https://www.compassion.com/multimedia/2018-annual-report-compassion-international.pdf`

Cox, David R. 1972. "Survival Analysis for the Unemployment Duration." *Journal of the Royal Statistical Society, Series B (Methodological)* 34 (2): 187-220.

Dunn, Olive Jean. 1961. "Multiple Comparisons Among Means." *Journal of the American Statistical Association* 56 (293): 52-64, March.

Gravois, John. 2004. "Why Do Adoptive Parents Prefer Girls?" *Slate*, January 14. `https://slate.com/news-and-politics/2004/01/why-do-adoptive-parents-prefer-girls.html`

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69 (1): 201-209, January.

Hansen, Mary Eschelbach, and Bradley A. Hansen. 2006. "The Economics of Adoption of Children from Foster Care." *Child Welfare* 85 (3): 559–583.

Harbaugh, William T., Ulrich Mayr, and Daniel Burghart. 2007. "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations." *Science* 316 (5831): 1622-1625, June.

Kuhn, Ben. 2014. "Decision Tree for Survival Analysis," November. `https://www.benkuhn.net/survival-trees`

List, John A. 2011. "The Market for Charitable Giving." *Journal of Economic Perspectives* 25 (2): 157–180, Spring.

Myerson, Roger B., Mark A. Satterthwaite. 1983. "Efficient Mechanisms for Bilateral Trading." *Journal of Economic Theory* 29 (1): 265–281, April.

Rodriguez, Germán. 2010. "Survival Models." *Lecture Notes on Generalized Linear Models*, September. `https://data.princeton.edu/wws509/notes`

Roth, Alvin E. 2002. "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics." *Econometrica* 70 (4): 1341–1378, July.

Wydick, Bruce, Paul Glewwe, and Laine Rutledge. 2013. "Does International Child Sponsorship Work? A Six-Country Study of Impacts on Adult Life Outcomes." *Journal Political Economy* 121 (2): 393–436.