SOCIUS

Special Collection: Fragile Families Challenge

AMERICAN SOCIOLOGICAL ASSOCIATION

Socius: Sociological Research for
a Dynamic World
Volume 5: 1–21
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2378023118814623
srd.sagepub.com
SAGE

Using LASSO to Assist Imputation and Predict Child Well-being

Diana Stanescu[1], Erik Wang[1], and Soichiro Yamauchi[2]

Abstract

This article documents an approach to predicting children's well-being using data from the Fragile Families and Child Wellbeing Study, which are representative of births in large U.S. cities. The authors use the least absolute shrinkage and selection operator (LASSO) to preprocess the data. They then apply the Amelia algorithm to impute missing data. Finally, they use LASSO again for prediction with the imputed data. The authors report the performance of this approach for six outcome variables. The approach achieves the best performance for the variable material hardship. The out-of-sample mean squared error of the authors' prediction is 0.019, the lowest among all submissions in the Fragile Families Challenge. The authors find that among variables with high predictive power, variables from mother surveys dominate. Furthermore, components of material hardship in the past strongly predict current material hardship.

Keywords

material hardship, prediction, LASSO, Fragile Families Challenge

In this article, we describe an approach assisted by the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996) to making predictions of material hardship and other measures of child well-being for children at age 15. Material hardship is a measure first developed by Mayer and Jencks (1989) of extreme poverty that aggregates positive responses to a set of survey questions. We use data originally from the Fragile Families and Child Wellbeing Study. To tackle the issues of missing data and variable selection, our approach consists of multiple steps: cleaning, preprocessing using LASSO, model-based imputation, and prediction using LASSO.

We apply this approach to predict material hardship, along with five other outcomes concerning children performance and welfare: grade point average (GPA), grit, job training, eviction, and layoff. We submit our results to the Fragile Families Challenge (FFC). The FFC is a mass collaborative effort with the goal of producing and facilitating research and policy ramifications aimed at addressing the challenge of fragile families in the United States. It invites scholars to make predictions of the six aforementioned outcomes using data from the Fragile Families and Child Wellbeing Study. The study produces data representative of births in large U.S. cities between 1998 and 2000. These data are based on mother and father interviews conducted at children's birth and at years 1, 3, 5, and 9.[1] It therefore has many advantages over surveys of a similar kind, chief among which is an oversample of nonmarital births (3:1) for which interviews were conducted with both mothers and fathers, thus obtaining rich information about them (Reichman et al. 2001). The lessons learned from these prediction exercises will make an important step toward accomplishing the FFC mission.[2]

The rest of this article is organized as follows. First we introduce LASSO as our main method. We then document our procedures of data cleaning, preprocessing, imputation, and prediction. Next we report the performance of our approach. Finally, we discuss the results by highlighting the importance of predictors from mother surveys and components of material hardship measured in the past.

LASSO as the Main Method

The use of LASSO underpins our strategy. In our approach, LASSO is used twice: first to preprocess the data and then to

[1] We refer to mother interviews as mother surveys and to father interviews as father surveys.

[2] For a complete description of the FFC and the data used in the FFC, as well as the six outcome variables, please refer to the introductory article in this special collection (Salganik et al. 2019).

[1] Princeton University, Princeton, NJ, USA
[2] Harvard University, Cambridge, MA, USA

Corresponding Author:
Erik Wang, Princeton University, Department of Politics, 025 Corwin Hall, Princeton, NJ 08544, USA.
Email: haixiaow@princeton.edu

train prediction models. LASSO handles high-dimensional data (i.e., the number of covariates can be larger than that of units) well because its penalization shrinks tiny coefficients to exactly zero. Selecting variables by zeroing out coefficients also makes postestimation analysis easier, as the number of covariates becomes much smaller, which is advantageous for preprocessing the high-dimensional FFC data set. In addition, LASSO helps avoid overfitting to the training data via regularization. This feature is helpful for building prediction models.

Given the training data $\{Y_i, X_i\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$, the LASSO estimate is defined so that it minimizes the squared loss with L_1 norm penalty, $\|\beta\|_1 = \sum_j |\beta_j|$. Formally, estimates for the LASSO are given by

$$\hat{\beta}_{\text{lasso}} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \|\beta\|_1 \right\}, \tag{1}$$

where $\lambda > 0$ is a tuning parameter that is chosen by cross-validation. The second term, $\lambda\|\beta\|_1$, works as a regularizer that encourages smaller β. Intuitively, when minimizing, a large λ would induce a smaller magnitude for β.

Here, variables are standardized to have zero mean and unit variance so that regularization on coefficients is not affected by the original scale of input variables and the intercept can be omitted from equation 1. One property of LASSO is that the estimated coefficient can be exactly zero (i.e., it can achieve variable selection). For a new input X_{n+1}, prediction is made by $\hat{Y}_{n+1} = X_{n+1}^{\mathrm{T}}\hat{\beta}_{\text{lasso}}$.

For binary outcomes, we use logistic regression with L_1 penalty. The estimates are given by

$$\hat{\beta}_{\text{logit-lasso}} \in \arg\min_{\beta} \left\{ -\left[\begin{array}{l} \frac{1}{n} \sum_{i=1}^{n} Y_i X_i^{\mathrm{T}}\beta \\ -\log(1+\exp(X_i^{\mathrm{T}}\beta)) \end{array} \right] + \lambda\|\beta\|_1 \right\}, \tag{2}$$

which corresponds to minimizing the negative log likelihood of the model with L_1 regularization.

We predict probabilities, instead of classes, for binary outcomes, as the FFC recommends.

Procedures of Data Preprocessing and Prediction

This section details our procedures of data cleaning, preprocessing, imputation, and prediction.[3]

Step 1: Cleaning

We immediately drop any variable with more than 60 percent of observations assigned NA (not applicable; meaning that values are missing) or negative values. In this dataset, negative values indicate different types of missingness. An extremely high degree of missingness would prevent such variables from conveying useful information for prediction purposes. We treat categorical variables as ordinal variables and apply the above cleaning rules. This procedure reduces the number of potential covariates from 12,942 to 4,207. We further exclude variables that either indicate the date of the survey only or have standard deviations less than 0.01. This step leaves us with 4,187 variables.[4]

Step 2: Preprocessing with LASSO to Assist Imputation

We want to identify a small set of covariates from these 4,187 variables. Missing values in this smaller set would be imputed with Amelia, a model-based imputation algorithm proposed by King et al. (2001).[5] To arrive at these covariates, we first mean-impute the covariates and use LASSO. We use LASSO here not to make immediate predictions but to determine this small set of variables for further use. To the best of our knowledge, there have not been any prior studies that used LASSO as a preprocessing tool in preparation for further imputation using model-based methods.

We regress the six outcomes separately on mean-imputed covariates in the FFC using LASSO.[6] For each of the six sets of results, we drop the covariates with coefficients of size zero. Then we take the union over the six sets of remaining variables. This procedure leaves us with 339 covariates, listed in Table A9 in the Appendix.

Step 3: Model-based Imputation with Amelia

We identify these 339 covariates (obtained with LASSO) in the original (i.e., before mean imputation) data set. We run a model-based imputation algorithm, Amelia, on these variables from the original data set so that they will enter our final prediction process with their missing values imputed in a principled manner. Amelia jointly models variables with multivariate normal distribution. The expectation-maximization algorithm is used to estimate the model by iterating between the model parameters, mean and covariance matrix, and missing values until convergence. We use model-based imputation here because we believe covariates are correlated with one another, and hence missing values are expected to have more accurate imputation by Amelia, which fully exploits the correlation structure of covariates.

Tables 1 and 2 summarizes how many covariates survived after each step in the cleaning, preprocessing and imputation stages.

After data cleaning and imputation for covariates, we also impute the outcome variables. We create an outcome matrix with

[3] All analyses are done in R version 3.4.3 (R Core Team 2017).

[4] Sixty percent missingness and 0.01 standard deviation cutoffs are arbitrary choices made without further sensitivity checks or consultation with existing studies.

[5] R package Amelia version 1.7.4 is used for the analysis (Honaker, King, and Blackwell 2011).

[6] R package glmnet version 2.0.13 is used to fit LASSO (Simon et al. 2011). Tuning parameters are selected by fivefold cross-validations.

Table 1. Number of Predictors Remaining after Each Data Preprocessing and Imputation Step.

Step		Variables Selected by Screening					
0	Original	12,942					
1	Remove missing ≥60 percent	4,207					
2	Remove variables with SD < 0.01	4,187					
3	LASSO (union)	339					
4	Imputation	339					
		Variables selected by LASSO for each outcome					
		Material hardship	GPA	Grit	Eviction	Layoff	Job training
5	LASSO	72	66	190	106	75	64

Note: Steps 0 through 4 correspond to the variable preprocessing and imputation stages, and step 5 corresponds to the prediction stage. GPA = grade point average; LASSO = least absolute shrinkage and selection operator.

Table 2. The Two Rows Detail the Number of NA Observations Remaining after the Outcome Variables Were Imputed (Total 2,121 Units).

	Number of NA Observations per Outcome					
	Material Hardship	GPA	Grit	Eviction	Layoff	Job Training
Original data	662	956	703	662	844	660
After imputation	655	655	655	655	655	655

Note: GPA = grade point average; NA = observations whose values are missing.

columns corresponding to each outcome variable and impute missing cells using Amelia. Outcomes for the same individual can be highly correlated. Information borrowed across outcomes should therefore improve the prediction of each outcome.[7] Tables 1 and 2 document the results of outcome imputation. Figure A1 in the Appendix shows correlations among outcome variables after imputation. Figure A2 displays distributions of imputed versus actual data among the six outcomes.

Step 4: Using LASSO (Again) to Predict Six Outcomes

After these three steps, we train prediction models with LASSO for each outcome using the R package glmnet (Simon et al. 2011). Binomial link (equation 2) is used for binary outcomes (eviction, layoff, and job training), and the linear model (equation 1) is used for GPA, grit, and material hardship. We choose tuning parameters by fivefold cross-validation for each outcome separately and select values that minimize mean squared error (MSE).

Results

The first row of Figure 1 displays the densities of out-of-sample predictions, in-sample fitted values, and in-sample training data for continuous outcomes. The second row shows separation plots (Greenhill, Ward, and Sacks 2011) for binary outcomes.

Table 3 reports MSEs of predictions.[8] "Final model" refers to results obtained using our approach described in this article. Each MSE in the "winning model" refers to the MSE obtained by the team that won the FFC for that corresponding variable. All other models come from post-FFC analysis. In these models, we replicate our analysis (1) using the sample mean of the imputed outcomes in testing data as predicted values for all testing units ("null model"), (2) skipping the Amelia imputation step and instead using mean imputation for all missing values ("mean imputation"), and restricting the covariates to (3) mother survey items only

[7] When working on this project, we thought that we should not use covariate information when imputing the outcome, because we wanted to avoid contamination. Our intuition was that the covariates that contributed a lot to imputation would also receive higher coefficients in the variable selection using LASSO, but these higher coefficients were induced by construction. After more careful consideration, we realized that this intuition might not necessarily be correct. We thus refrain from advocating this particular choice of imputing outcome using only information about other outcomes but not covariates in this article. When imputing both outcome and covariates, we set the number of imputed datasets by Amelia, M, to 5, and choose the third one. It was an arbitrary decision of ours regarding the size of M and which one(s) to use.

[8] For complete out-of-sample MSEs for all six outcomes in both leaderboard and holdout data, refer to Table A8 in the Appendix.

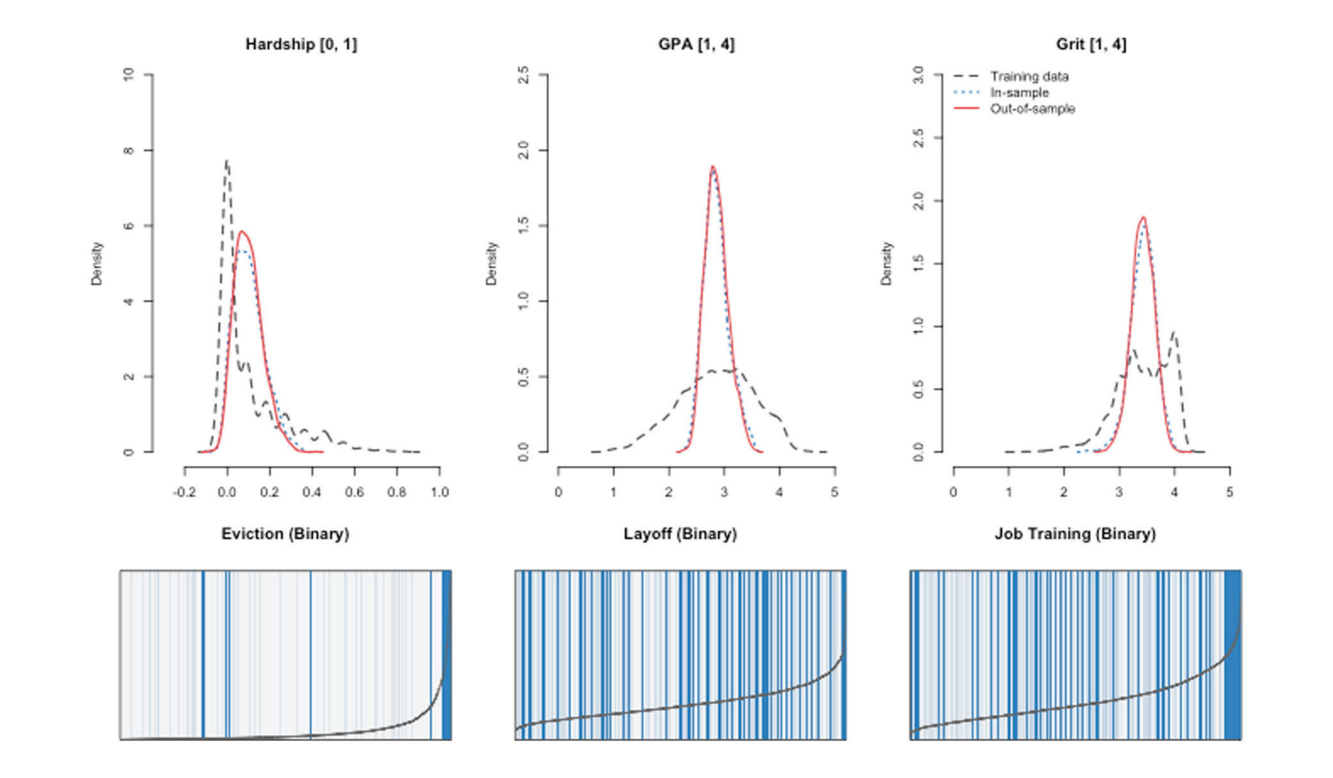


Figure 1. Density plot (first row) and separation plot (second row) for predicted outcomes. First row: red solid lines represent out-of-sample predicted outcome. Blue dotted lines represent in-sample fitted values. Black dashed lines are densities of outcomes in training data. Second row: separation plot for binary outcomes. Predicted probabilities for the training set are sorted according to the predicted probability from the left (minimum) to the right (maximum) and then colored by the actual outcome. The blue vertical lines occur at points where the observation takes the value 1 rather than 0. The superimposed black curve represents the predicted probabilities for the testing data set.

Table 3. Results of Predictions (MSE on Holdout Data).

	Hardship	Grit	GPA	Eviction	Layoff	Job Training
Final model	**0.019**	0.253	0.361	0.059	0.167	0.181
Winning model	0.019	0.238	0.344	0.052	0.162	0.176
Null model	0.025	0.253	0.426	0.055	0.167	0.185
Mean imputation	0.020	0.257	0.357	0.057	0.178	0.185
Mother only	0.019	0.249	0.389	0.055	0.164	0.175
Father only	0.024	0.253	0.395	0.054	0.166	0.185

Note: "Final model" refers to results obtained using the approach described in this article. Each mean squared error (MSE) in the "winning model" refers to the MSE obtained by the team that won the Fragile Families Challenge for the corresponding variable. All other models come from postchallenge analysis. In these models, we replicate our analysis (1) using the sample mean of the imputed outcomes in testing data as predicted values for all testing units, (2) skipping the Amelia imputation step and instead using mean imputation for all missing values, and restricting the covariates to (3) mother survey items only and (4) father survey items only. GPA = grade point average.

("mother model") and (4) father survey items only ("father model").[9,10]

The out-of-sample prediction of material hardship using our approach achieves an MSE of 0.019, the lowest among all FFC submissions for this variable. With respect to rankings, our approach was also competitive for the following outcomes: GPA and job training. Among 163 submissions, the rankings are 30 for GPA and 30 for job training but below 100 for the other three outcomes.

[9] We thank the editor and the FFC team for helping us obtain the post-FFC analysis results.

[10] Mother model and father model follow the same procedure up to the Amelia imputation step.

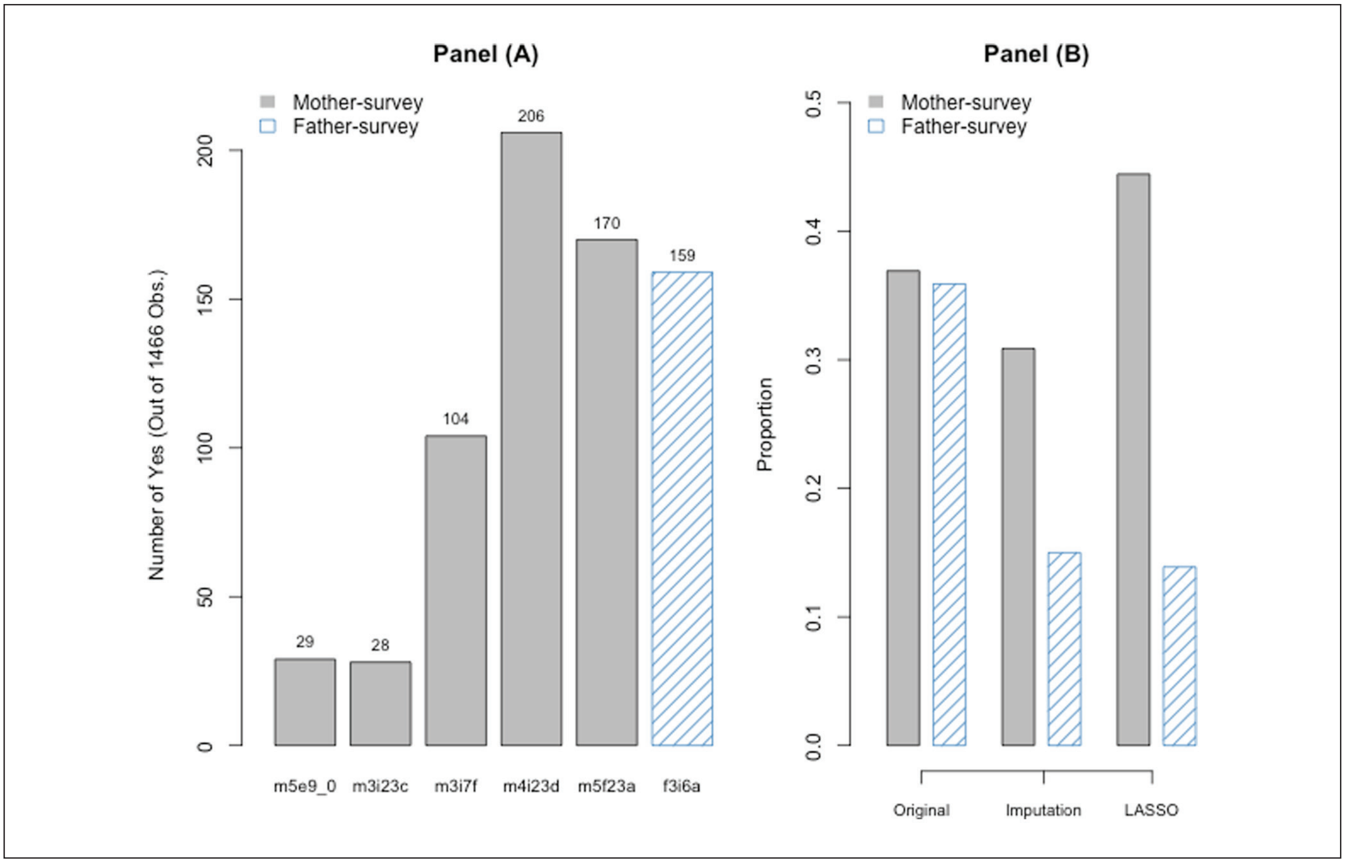


Figure 2. (A) Count plot of "yes" answers for each variable. Variable names refer to the following: m5e9_0, only person from whom the child seeks help; m3i23c, evicted from home; m3i7f, helped by employment office; m4i23d, could not pay mortgage; m5f23a, received free food or meals; and f3i6a, telephone disconnected. (B) Proportion plot of mother-survey (gray) and father-survey (blue) variables in the data set at each stage of preprocessing and prediction. The leftmost two bars correspond to the original data set; the middle two bars correspond to the imputed data set after removing missing variables, preprocessing with the least absolute shrinkage and selection operator (LASSO), and imputing with Amelia; and the rightmost two bars correspond to selected variables by the LASSO in predicting material hardship. We calculate the proportion by counting the variable names whose prefixes begin with the letter *m* for mother-survey variables, or *f* (but not "ffcc") for father-survey variables. Proportions do not sum to 1, because the data set contains answers to surveys not directed at the mother or father.

Regarding our models, we note that our approach, in general, performs better than the "null model" and the "mean imputation" model. However, the "mean imputation" model still performs comparably well, suggesting that Amelia imputation might not have improved the prediction as much as expected. Results for variables from mother surveys compared with those from father surveys are discussed next.

Discussion

In this section, we focus our discussion on material hardship. LASSO selects 72 variables for final prediction, listed with coefficients in Appendix Table A1. Our results reflect that variables from mother surveys are more helpful than those from father surveys in predicting material hardship.

Below we rank the selected variables in terms of the size of their rescaled coefficients. Because glmnet returns coefficients on the original scale, we manually rescale the coefficients, which approximates the standardized coefficients. Let $\hat{\beta}$ be the output from glmnet. The rescaled coefficient for variable j is given by $\hat{\beta}_j^* = \hat{\beta}_j \cdot \hat{\sigma}_Y / \hat{\sigma}_{X_j}$, where $\hat{\sigma}_Y$ and $\hat{\sigma}_{X_j}$ are the estimated standard deviations for Y and X_j, respectively. When reporting the rescaled coefficients, we drop the $\hat{\sigma}_Y$ term because this is constant across variables. We acknowledge that the ranking of variables here is simply a heuristic that aids substantive interpretation of the model, and we are not making any formal inference on these rankings.

The variable with the largest coefficient magnitude is whether the school instruction language is an Asian language for the child in year 5 (t5e7_3). However, in the original data, there are only 2 people answering "yes" but 2,004 people answering "no" to the survey question, with 52.7 percent of the observations missing. The variation that drives our prediction mostly comes from imputation. In addition, when rescaling coefficients, we divide glmnet estimates by empirical standard deviations. This procedure mechanically produces large (rescaled) coefficients when the original variable has small variation. Figure 2A shows the variables with the second largest to the sixth largest coefficient magnitudes. They are (1) whether the child in year 5 asks no one for help or advice other than the mother (m5e9_0), (2) whether the mother in year 3 was evicted from home in the past year (m3i23c), (3) whether the mother in year 3 was helped by an employment office since the child's first birthday (m3i7f), (4) whether the mother in year 5 could not complete mortgage payments for the past 12 months because there was not enough money (m4i23d), and (5) whether the mother in year 5 received free food or meals over the past 12 months (m5f23a).

We draw two main conclusions. First, these variables share one common characteristic: they are from mother surveys. The highest predictive variable from father surveys is whether the father in year 3 noticed the telephone disconnected in the past 12 months (f3i6a). This variable ranks 11th among our 72 selected variables. We further verify the performance gap between mother- and father-survey items in a post hoc analysis that compares the prediction results obtained using just the variables from mother surveys against those obtained using just the variables from father surveys. As shown in Table 3, the MSE for material hardship is 0.019 for the former (which is as low as that obtained using the LASSO-assisted approach described in this article) and 0.024 for the latter. Notably, using just the mother-only model will lead to better prediction results than those obtained using our approach in this article for four of six outcomes.

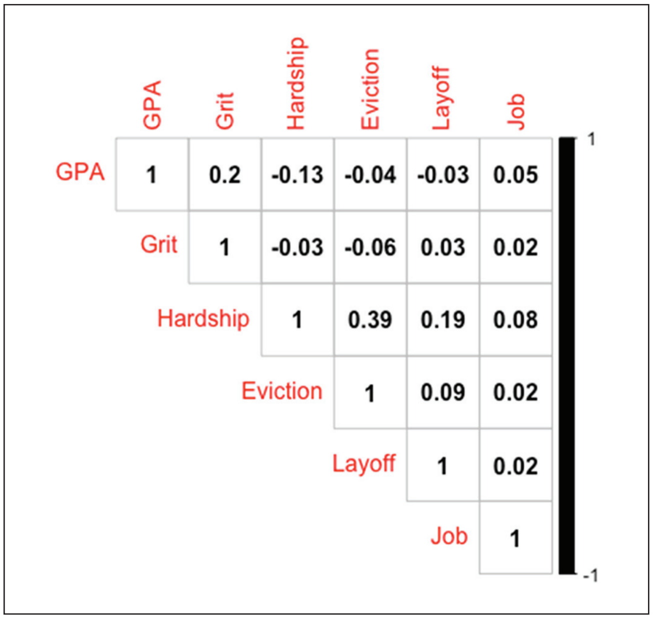
One may attribute the performance gap between mother-survey variables and father-survey variables in step 4 of our LASSO-assisted approach to various factors. For one, variables from father surveys are more likely to have substantial amount of missing values and so are less likely to survive in the initial stages of data cleaning, in which we delete variables according to the 60 percent cutoff rule described earlier. Figure 2B shows that items from father surveys start to have much lower proportions than those from mother surveys at the imputation stage. Yet the difference in proportions further increases after LASSO, indicating that the performance gap is more than an artifact of data cleaning. Moreover, it may be interesting in itself that father survey variables are more likely to suffer from missing values than variables from mother surveys. We want to acknowledge, however, that prediction is completely different from causal inference. Whether the importance of mother-survey predictors over those from father surveys indicates anything causal about the substantive importance of mother's role in family welfare, childcare, or child's education goes beyond the scope of this article.

Second, our results suggest that past outcomes may effectively predict current outcomes in panel data. Questions from which variables 2, 4, and 5 are constructed, as well as the top-ranked variables from father surveys, were similar to those asked in the year 15 primary caregiver survey that would in turn form 4 of 11 components of material hardship. Social scientists have long used past outcomes to predict future outcomes. Hegre et al. (2013) is a prominent example showing that recent history of a country's armed conflict is a robustly effective predictor of the country's future conflict. Whether past material hardship necessarily causes future material hardship or simply reflects some unobserved underlying causes that are correlated across time may be a subject of future research.

Appendix to Predicting Material Hardship: Using LASSO to Assist Imputation and Select Variables

Additional Figures

Figure A1 shows a correlation matrix of outcome variables.



	GPA	Grit	Hardship	Eviction	Layoff	Job
GPA	1	0.2	-0.13	-0.04	-0.03	0.05
Grit		1	-0.03	-0.06	0.03	0.02
Hardship			1	0.39	0.19	0.08
Eviction				1	0.09	0.02
Layoff					1	0.02
Job						1

Figure A1. Correlation plot of outcome variables. Correlations are computed on the basis of postimputation data.

Additional Tables

Tables A1 through A6 show the summary of the variables selected out of the prediction model for each outcome, as specified in the captions. The first column shows the variable names per the original data set and codebooks. The second and third columns present regression coefficients from LASSO. Coefficients in the second column are in original scale, while those in the third column are standardized. Columns 4 to 7 show the summary statistics for each variable.

Table A7 shows common variables selected as predictive across models.

Table A8 shows out-of-sample results of predictions for six outcomes. Reported numbers are MSEs. The first two rows

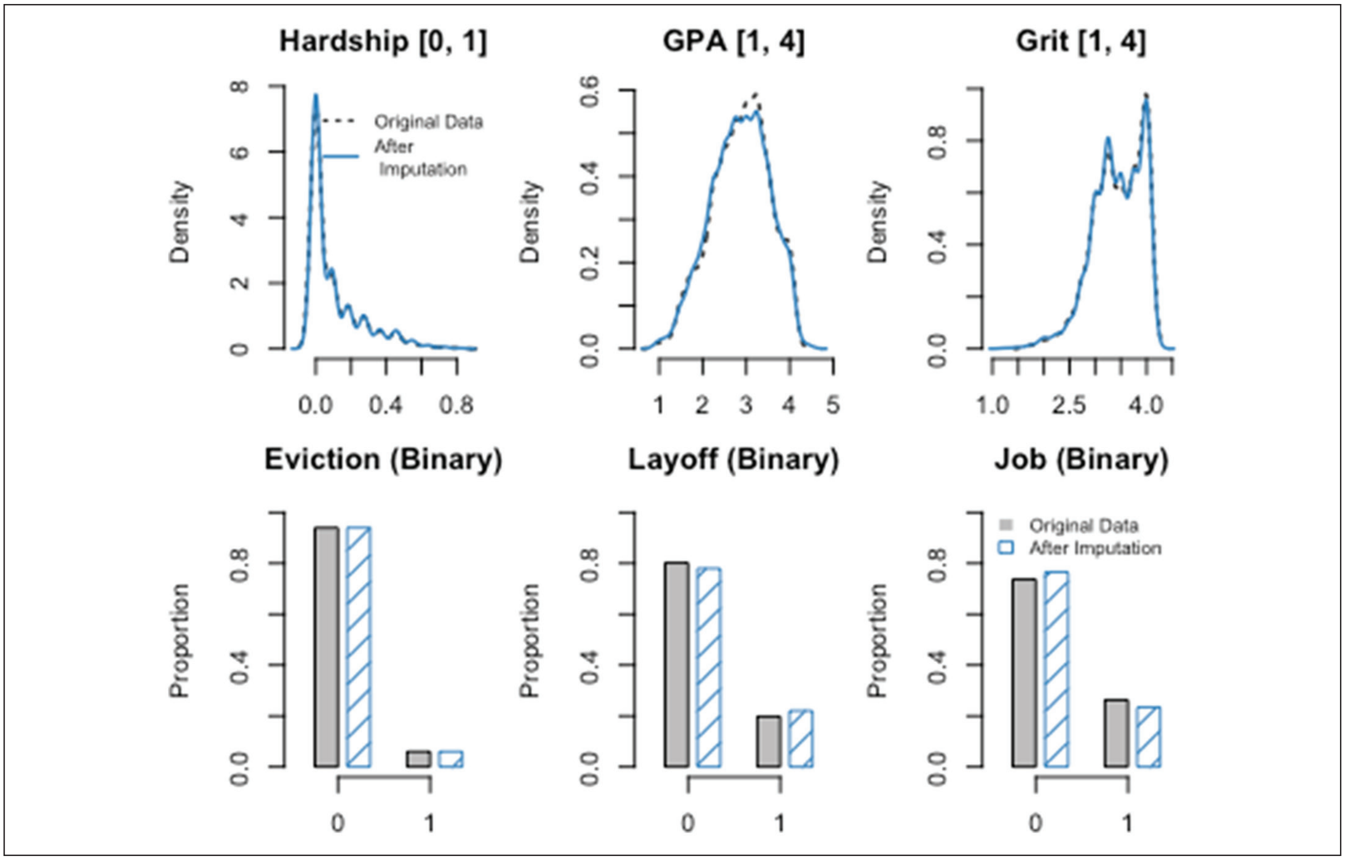


Figure A2. Displays distribution of outcome variables before and after imputation using the Amelia algorithm. The top row consists of three density plots of continuous outcomes, and the bottom row consists of three bar plots for binary outcomes. The numerical range next to the variable name indicates the support of each variable (e.g., GPA takes values between 1 and 4). Hardship here refers to material hardship.

Table A1. Summary of Variables Selected out of the Prediction Model for Material Hardship.

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
t5e7_3	−0.019	−0.674	1.000	2.077	2.000	2.000
m5e9_0	0.029	0.219	−0.438	1.000	0.018	0.000
m3i23c	−0.021	−0.149	1.000	2.469	1.979	2.000
m3i7f	−0.027	−0.102	1.000	2.682	1.923	2.000
m4i23d	−0.031	−0.091	0.640	3.130	1.865	2.000
m5f23a	−0.028	−0.086	0.850	2.909	1.880	2.000
p5l12b	−0.014	−0.076	1.000	2.697	1.966	2.000
m3k27a	−0.013	−0.060	1.000	2.681	1.953	2.000
m5f23e	−0.025	−0.055	0.309	3.271	1.700	2.000
hv3p6_e	0.012	0.055	−0.684	1.000	0.049	0.000
m4k26a	−0.013	−0.051	1.000	2.710	1.933	2.000
f3i6a	−0.017	−0.050	0.629	2.892	1.853	2.000
m5f23c	−0.019	−0.049	0.427	2.875	1.805	2.000
m4i23n	−0.019	−0.049	0.805	3.063	1.813	2.000
m5i14a3	−0.014	−0.044	0.889	2.907	1.881	2.000
m2h18	−0.017	−0.041	0.290	3.080	1.714	2.000
m5f23k	−0.016	−0.039	0.546	2.990	1.794	2.000
f5g28	−0.012	−0.037	0.837	2.906	1.882	2.000
m3i23d	−0.014	−0.033	0.705	3.056	1.777	2.000

(continued)

Table A1. (continued)

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
m3b24	−0.011	−0.032	0.723	2.960	1.870	2.000
hv4l59	0.010	0.031	−0.987	2.000	0.102	0.000
f5i14b4	0.015	0.031	0.099	3.506	1.636	1.841
m5f23g	−0.014	−0.030	0.169	3.323	1.703	2.000
hv3s1_1	0.005	0.028	−0.485	1.000	0.026	0.000
hv3d2	0.013	0.027	−1.123	1.651	0.306	0.000
p5q3ag	0.008	0.026	0.119	3.000	1.081	1.000
m3l6a	0.004	0.024	0.301	2.000	1.056	1.000
m2h9a1	−0.010	−0.023	0.520	3.057	1.751	2.000
p5q3k	0.010	0.023	−0.408	3.000	1.207	1.000
p5q3bl	0.008	0.022	−0.103	3.000	1.128	1.000
p5h15	−0.009	−0.020	0.331	2.982	1.726	2.000
f2a7d	−0.009	−0.019	0.120	2.890	1.567	1.761
m2g5	−0.009	−0.018	0.029	3.082	1.589	2.000
hv4k9	0.007	0.017	−1.018	2.106	0.744	1.000
p5q3bo	0.007	0.017	−0.038	3.000	1.206	1.000
f2g13	−0.011	−0.016	−0.741	3.662	1.561	1.451
m5g0	0.011	0.015	−0.241	4.161	1.734	2.000
p5q3bn	0.007	0.015	−0.288	3.000	1.203	1.000
cm4marp	0.003	0.013	−0.555	1.000	0.039	0.000
m4c38	−0.003	−0.012	1.000	2.765	1.923	2.000
m3i23e	−0.005	−0.011	0.619	3.253	1.770	2.000
f4l6	0.004	0.010	0.015	2.634	1.207	1.000
m4b2	0.007	0.009	−0.761	5.000	1.529	1.000
m4i9	−0.003	−0.007	0.717	3.235	1.816	2.000
m2h19h	−0.002	−0.007	0.937	2.834	1.875	2.000
m3k3c	−0.001	−0.006	1.000	2.640	1.939	2.000
m5e6	0.002	0.004	−0.005	2.850	1.417	1.141
m5g16b	0.003	0.004	0.669	5.591	3.400	4.000
p5q3by	0.001	0.004	0.152	3.000	1.104	1.000
hv4f1f	−0.004	−0.004	−0.213	8.296	3.632	4.000
k5f1	0.002	0.003	8.032	12.466	9.995	9.814
f4b4b2	−0.002	−0.003	−1.793	3.315	0.699	0.759
hv3m2b	0.003	0.003	−1.787	3.509	0.818	1.000
cm5edu	0.003	0.003	−1.086	6.170	2.513	2.936
m5g2c	−0.001	−0.003	0.490	3.156	1.780	2.000
p5q2d	−0.003	−0.003	1.000	12.038	7.776	8.000
m5e8_5	−0.001	−0.003	−1.063	1.989	0.545	0.684
m4j0	0.001	0.002	−0.248	4.036	1.685	2.000
p5q3dk	−0.001	−0.002	0.061	4.241	2.282	2.000
m5g1	0.002	0.002	−0.966	5.401	2.427	2.137
p5j2j	0.002	0.002	−2.103	5.773	2.161	2.000
cf4povcab	−0.002	−0.001	−1.519	8.159	3.313	3.151
f5k3b	0.001	0.001	−1.222	4.284	1.233	1.171
p5q3dd	0.000	−0.001	0.312	4.576	2.434	2.592
k5e1d	−0.001	−0.001	−0.073	7.176	3.380	4.000
hv3j11	0.001	0.001	−5.206	9.047	2.335	2.145
hv3j7	0.001	0.000	−3.765	9.047	2.709	3.000
p5i31h	−0.001	0.000	−1.969	8.903	3.356	3.643
f3c3g	0.000	0.000	−2.749	11.868	4.951	5.000
p5j11	0.000	0.000	−29.742	101.000	1.670	1.000
p5q1j	0.000	0.000	−2.471	13.259	5.003	5.000
f3i4	0.000	0.000	−465.165	1,605.001	495.466	484.229

Note: The first column shows the variable names as in the original data set and codebooks. The second and third columns present regression coefficients from the least absolute shrinkage and selection operator. Coefficients in the second column are in original scale, while those in the third column are standardized. Columns 4 to 7 show the summary statistics for each variable.

Table A2. Summary of Variables Selected out of the Prediction Model for Grade Point Average.

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
hv4r10a_3	−0.146	−2.508	−0.166	1.000	0.002	0.000
o5d1_6	−0.095	−1.108	1.000	2.263	1.991	2.000
m5e9_0	0.090	0.679	−0.438	1.000	0.018	0.000
hv3s1_1	0.077	0.443	−0.485	1.000	0.026	0.000
p5l13f	−0.050	−0.159	0.980	3.036	1.890	2.000
p5q3bb8	−0.024	−0.158	0.565	3.000	1.016	1.000
f2h5a	0.028	0.112	1.000	2.742	1.935	2.000
t5b3e	−0.057	−0.100	−1.148	4.000	1.301	1.000
m3i23e	0.038	0.090	0.619	3.253	1.770	2.000
p5i14	0.036	0.077	−0.026	2.745	1.301	1.000
o5f6	−0.039	−0.074	1.000	6.415	4.808	5.000
m4f2e2	−0.031	−0.069	−0.387	3.044	1.303	1.000
t5b1f	0.056	0.064	−0.023	5.822	2.980	3.000
hv3c5	0.029	0.058	−1.435	2.320	0.496	0.485
hv4l59	−0.018	−0.055	−0.987	2.000	0.102	0.000
m5j2	0.025	0.052	−0.272	3.223	1.594	1.913
p5q3u	−0.025	−0.045	−0.212	3.332	1.504	1.449
o5f4	−0.024	−0.045	1.000	6.427	4.796	5.000
p5i20c	0.012	0.041	−0.048	2.041	1.100	1.000
hv4l47	−0.016	−0.041	−1.116	2.000	0.142	0.000
m5f23c	0.016	0.041	0.427	2.875	1.805	2.000
m5i3c	−0.011	−0.039	1.000	2.899	1.924	2.000
f1b20	−0.019	−0.038	−0.355	3.187	1.358	1.124
hv4sex_child	0.018	0.036	−0.156	3.011	1.480	1.441
t5c16	0.024	0.034	0.570	5.716	3.076	3.000
m3i8a3	0.008	0.033	1.000	2.944	1.934	2.000
m3b5	−0.015	−0.030	−0.189	2.803	1.475	1.304
hv4b9	0.014	0.029	−0.680	2.290	0.628	0.943
hv3m2c	−0.020	−0.027	−1.594	3.240	1.058	1.000
m5g19	0.019	0.025	−1.678	4.000	0.803	1.000
f4i23d	−0.007	−0.023	0.586	2.812	1.877	2.000
k5g2h	−0.022	−0.023	−2.514	4.370	0.814	0.776
f4h1q	0.011	0.016	−1.562	5.000	1.373	1.000
m4i9	0.006	0.015	0.717	3.235	1.816	2.000
p5m1	−0.017	−0.013	−1.355	7.320	3.317	3.737
m5b30	0.006	0.013	0.285	3.494	1.725	2.000
m1i1	0.023	0.013	1.000	9.000	4.683	4.000
t5b1u	0.011	0.012	−0.707	5.260	2.375	2.146
f2k12	0.005	0.012	−0.308	2.838	1.301	1.000
t5b1w	0.008	0.010	0.132	5.427	2.945	3.000
cm2povco	0.014	0.009	−2.858	6.658	1.727	1.314
hv3c8	0.009	0.008	−1.376	5.377	1.815	1.990
p5i23	0.011	0.008	−1.799	7.239	3.234	3.084
cm4marp	0.002	0.008	−0.555	1.000	0.039	0.000
f5k14b	0.008	0.007	−1.956	6.060	1.908	2.000
m3k22	−0.009	−0.007	−2.484	30.000	1.568	1.000
t5b1d	0.005	0.006	0.221	5.689	2.853	3.000
p5i31h	0.005	0.003	−1.969	8.903	3.356	3.643
p5m2e	−0.003	−0.003	−2.001	4.933	1.794	1.703
f3k12	−0.006	−0.002	96.773	114.330	105.492	105.000
f4i0n2	0.002	0.002	−0.666	5.538	2.041	2.000
hv3g1f	0.002	0.002	−0.684	8.017	3.635	4.000
m1i3	0.003	0.002	−0.870	9.196	4.617	4.000
cf5povco	0.003	0.002	−3.569	8.624	2.349	2.085

(continued)

Table A2. (continued)

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
hv4d15c	−0.003	−0.002	−2.936	7.712	2.628	2.632
k5a3c	0.001	0.001	−1.840	5.302	1.872	2.000
k5b1b	−0.001	−0.001	−2.287	5.789	1.594	1.745
m3l3	0.000	−0.001	−0.698	2.808	1.371	1.000
f5k7	0.002	0.001	−10.113	30.000	2.816	2.000
hv3m2b	0.000	0.000	−1.787	3.509	0.818	1.000
hv5_ppvtpr	0.001	0.000	−49.223	137.714	36.153	32.000
hv5_wj10pr	0.001	0.000	−34.456	156.657	47.855	47.000
f2g1a	0.001	0.000	−107.895	180.060	19.950	1.000
p5j10	0.000	0.000	−125.486	263.493	59.553	46.494
f3i4	0.000	0.000	−465.165	1,605.001	495.466	484.229
f5i13	0.000	0.000	−97,875.725	145,822.887	17,753.259	7,873.140

Note: The first column shows the variable names as in the original data set and codebooks. The second and third columns present regression coefficients from the least absolute shrinkage and selection operator. Coefficients in the second column are in original scale, while those in the third column are standardized. Columns 4 to 7 show the summary statistics for each variable.

Table A3. Summary of Variables Selected out of the Prediction Model for Grit.

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
t5e7_3	0.406	14.121	1.000	2.077	2.000	2.000
hv4r10a_3	−0.798	−13.681	−0.166	1.000	0.002	0.000
m1c1d	−0.173	−1.447	0.878	3.000	1.012	1.000
m2d2	−0.135	−0.958	0.565	2.000	1.022	1.000
m3i23c	0.124	0.866	1.000	2.469	1.979	2.000
cm4fdiff	0.035	0.711	−0.136	1.000	0.003	0.000
hv3v6b	−0.114	−0.526	−0.582	2.000	0.036	0.000
f5a7	−0.110	−0.496	1.000	2.776	1.936	2.000
f5g23	0.077	0.348	1.000	2.776	1.944	2.000
m3i8a3	0.083	0.338	1.000	2.944	1.934	2.000
p5l12b	0.060	0.335	1.000	2.697	1.966	2.000
m2d3b5	−0.049	−0.334	1.000	2.390	1.975	2.000
hv3a27d	0.053	0.298	0.000	1.571	0.967	1.000
p5q3bb8	−0.040	−0.267	0.565	3.000	1.016	1.000
m5e8_7	0.071	0.267	−0.757	1.002	0.076	0.000
m5e9_0	0.032	0.244	−0.438	1.000	0.018	0.000
m2f5	−0.065	−0.222	1.000	2.825	1.908	2.000
f5c1f	0.047	0.186	0.291	3.000	1.068	1.000
p5q3bl	0.066	0.173	−0.103	3.000	1.128	1.000
cm4marp	0.033	0.172	−0.555	1.000	0.039	0.000
hv3j19	−0.051	−0.146	−1.233	7.000	0.023	0.000
m3b5	−0.072	−0.145	−0.189	2.803	1.475	1.304
f3i6h	0.034	0.139	1.000	2.766	1.936	2.000
f3i23e	0.053	0.131	0.468	3.041	1.801	2.000
p5h15	−0.057	−0.129	0.331	2.982	1.726	2.000
t5a9p	0.025	0.124	1.000	2.628	1.959	2.000
p5l17d	0.049	0.118	0.520	3.055	1.777	2.000
cm1bsex	−0.059	−0.118	0.951	2.000	1.476	1.000
hv3s4	0.030	0.117	0.002	3.000	1.044	1.000
hv3r12	0.038	0.113	−1.013	1.139	0.120	0.000
hv4l47	−0.043	−0.110	−1.116	2.000	0.142	0.000
p5l15	−0.040	−0.110	0.722	3.039	1.843	2.000
m5f23c	0.041	0.107	0.427	2.875	1.805	2.000

(continued)

Table A3. (continued)

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
f5g3	0.037	0.097	0.436	3.038	1.817	2.000
m5g24	0.038	0.097	0.457	2.909	1.810	2.000
f5g28	0.030	0.094	0.837	2.906	1.882	2.000
hv4d1a	−0.048	−0.091	0.691	4.654	2.745	3.000
hv3t1	−0.023	−0.086	0.000	1.818	0.927	1.000
f2k12	0.038	0.085	−0.308	2.838	1.301	1.000
hv4l29	0.027	0.083	−0.957	2.000	0.103	0.000
m4h3a	0.039	0.080	0.051	3.397	1.390	1.058
m5g28	0.028	0.076	0.230	2.909	1.815	2.000
hv3m44	−0.043	−0.076	−1.514	2.409	0.484	0.331
m3k27a	−0.015	−0.073	1.000	2.681	1.953	2.000
p5q3bn	0.032	0.070	−0.288	3.000	1.203	1.000
m3i25	−0.019	−0.068	1.000	2.896	1.914	2.000
t5c5	0.015	0.067	0.261	2.000	1.057	1.000
m3i18	0.024	0.064	0.601	2.945	1.803	2.000
p5i26	0.040	0.064	1.000	5.544	3.663	4.000
p5q3at	−0.023	−0.063	−0.006	3.000	1.122	1.000
m4b4b1	−0.036	−0.057	−1.368	2.430	0.493	0.000
t5e15b	−0.028	−0.054	−0.573	4.000	1.249	1.000
m5e8_5	−0.026	−0.053	−1.063	1.989	0.545	0.684
hv4l42	−0.026	−0.051	−1.567	2.114	0.253	0.000
m4f2e2	−0.022	−0.048	−0.387	3.044	1.303	1.000
p5q3cb	−0.024	−0.046	−0.435	3.000	1.290	1.000
f4b5	0.021	0.045	0.253	3.278	1.676	1.845
t5c15	−0.014	−0.045	0.893	3.032	1.883	2.000
o5f4	−0.023	−0.043	1.000	6.427	4.796	5.000
p5q3k	−0.019	−0.043	−0.408	3.000	1.207	1.000
m5g31	0.014	0.043	0.807	3.014	1.869	2.000
f5e9_4	0.021	0.041	−1.053	2.116	0.483	0.458
k5g1e	0.030	0.040	0.000	4.709	2.470	2.864
o5f6	−0.019	−0.037	1.000	6.415	4.808	5.000
f4i0n1	0.030	0.036	−1.127	4.483	1.863	2.000
f4b4b2	0.025	0.036	−1.793	3.315	0.699	0.759
m2g8	−0.017	−0.034	−0.295	3.311	1.554	1.639
p5j1	0.025	0.034	−0.434	4.410	2.221	2.000
m4k26a	−0.008	−0.033	1.000	2.710	1.933	2.000
hv4d2	0.014	0.033	−1.210	1.509	0.243	0.000
m5f23k	0.013	0.032	0.546	2.990	1.794	2.000
o5g7	−0.015	−0.030	0.092	3.232	1.555	1.727
k5g1c	0.027	0.030	−0.838	5.103	2.159	2.000
p5i18b	−0.027	−0.030	−1.737	6.000	0.866	1.000
k5g1b	0.025	0.029	−0.418	4.956	2.314	2.417
m1g4	−0.014	−0.029	1.000	4.323	3.772	4.000
m3k3c	−0.006	−0.027	1.000	2.640	1.939	2.000
m4b6c	−0.027	−0.027	0.126	6.330	3.236	3.973
f4j4	0.013	0.027	−0.248	3.341	1.421	1.300
f5k3b	0.020	0.026	−1.222	4.284	1.233	1.171
m2d2c	−0.018	−0.026	−0.274	4.000	1.422	1.000
m5j2	0.013	0.026	−0.272	3.223	1.594	1.913
m2h19h	0.009	0.025	0.937	2.834	1.875	2.000
p5i30a	0.013	0.024	0.059	3.247	1.633	2.000
m5g19	0.018	0.024	−1.678	4.000	0.803	1.000
m5i16c	−0.025	−0.024	−1.583	5.662	1.835	1.468
m3l6a	0.004	0.024	0.301	2.000	1.056	1.000

(continued)

Table A3. (continued)

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
t5c16	0.017	0.024	0.570	5.716	3.076	3.000
hv3m49	−0.016	−0.023	−1.510	2.724	0.696	0.970
m5f23b	−0.005	−0.022	1.000	2.658	1.933	2.000
f5i14b4	−0.010	−0.021	0.099	3.506	1.636	1.841
p5q3bb5	0.009	0.021	−0.036	3.000	1.152	1.000
t5a4	−0.006	−0.021	0.187	2.312	1.093	1.000
k5g2h	−0.020	−0.021	−2.514	4.370	0.814	0.776
m5g16b	0.018	0.020	0.669	5.591	3.400	4.000
hv4k9	0.009	0.020	−1.018	2.106	0.744	1.000
p5q3bt	−0.011	−0.020	−0.287	3.354	1.379	1.000
m2b9	0.007	0.019	0.703	2.730	1.842	2.000
o5f3	−0.015	−0.018	1.000	7.366	4.509	5.000
p5i14	0.008	0.018	−0.026	2.745	1.301	1.000
k5e2c	0.013	0.018	−2.183	4.000	0.264	0.000
t5b1d	0.015	0.018	0.221	5.689	2.853	3.000
o5a2	0.013	0.017	−1.015	4.445	1.635	1.635
f5k14b	0.019	0.017	−1.956	6.060	1.908	2.000
p5q3a	−0.008	−0.016	−0.310	3.006	1.322	1.000
t5a9o	0.003	0.016	1.000	2.578	1.955	2.000
t5b1o	0.014	0.016	−0.299	5.953	2.703	2.821
f4i0n5	−0.013	−0.016	0.659	5.947	3.186	3.000
f3i6a	0.005	0.016	0.629	2.892	1.853	2.000
t5b3e	−0.009	−0.015	−1.148	4.000	1.301	1.000
hv3m21	0.009	0.015	−1.191	2.922	0.586	0.628
m3l3	−0.007	−0.014	−0.698	2.808	1.371	1.000
hv3a11	0.003	0.014	−0.664	1.000	0.050	0.000
hv3k3f	0.014	0.014	−1.513	5.367	1.789	1.433
m5i3c	−0.004	−0.013	1.000	2.899	1.924	2.000
f4k3b	0.005	0.013	0.474	3.032	1.820	2.000
m5b22b	−0.016	−0.013	−1.816	6.664	2.647	2.730
k5g2f	−0.014	−0.012	−2.164	4.713	1.237	1.000
p5q3ag	−0.004	−0.012	0.119	3.000	1.081	1.000
cf4cohp	0.004	0.011	−0.886	1.233	0.117	0.000
m4i7f	0.005	0.011	0.281	3.236	1.751	2.000
p5j2e	0.011	0.011	−2.222	5.000	0.549	0.000
cm5md_case_lib	−0.004	−0.011	−0.918	1.353	0.170	0.000
k5a3c	0.011	0.010	−1.840	5.302	1.872	2.000
m5g0	−0.007	−0.010	−0.241	4.161	1.734	2.000
p5m2e	−0.009	−0.009	−2.001	4.933	1.794	1.703
cm5edu	−0.009	−0.009	−1.086	6.170	2.513	2.936
k5a1b	0.009	0.009	−1.681	5.260	2.193	2.633
hv4d15c	0.012	0.009	−2.936	7.712	2.628	2.632
p5i1j	0.009	0.009	1.000	8.042	4.432	5.000
k5g2d	−0.009	−0.008	−2.837	4.287	0.911	1.000
m5e1k	−0.012	−0.008	−1.371	7.899	2.888	3.000
k5b1b	−0.010	−0.008	−2.287	5.789	1.594	1.745
k5d1h	0.013	0.008	−3.549	6.144	1.387	1.000
f5k14a	−0.010	−0.008	−2.372	5.166	1.608	1.573
k5b2b	−0.009	−0.008	−2.688	5.044	1.007	1.000
p5j7a	0.007	0.008	0.521	6.753	3.466	4.000
cm5fevjail	−0.004	−0.007	−1.072	1.986	0.452	0.284
f2a7d	0.003	0.007	0.120	2.890	1.567	1.761
k5a2f	−0.006	−0.007	−0.905	4.905	1.895	2.000

(continued)

Table A3. (continued)

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
f2g13	0.004	0.007	−0.741	3.662	1.561	1.451
p5m1	−0.008	−0.006	−1.355	7.320	3.317	3.737
k5f1	−0.003	−0.006	8.032	12.466	9.995	9.814
p5k1e	−0.004	−0.006	−0.684	4.344	1.909	2.000
p5m2c	−0.005	−0.005	−1.171	5.296	1.806	1.693
k5e1c	0.007	0.005	−1.358	7.319	3.081	3.773
p5i33b	0.012	0.005	−1.116	15.751	6.810	8.000
p5i1f	0.006	0.005	−0.832	5.868	2.821	3.000
f4c7e	0.002	0.005	−0.172	3.076	1.370	1.224
t5b1u	0.004	0.005	−0.707	5.260	2.375	2.146
m5f23g	0.002	0.005	0.169	3.323	1.703	2.000
hv4d17	−0.006	−0.004	−2.354	6.721	2.136	1.924
cf5povco	0.008	0.004	−3.569	8.624	2.349	2.085
hv4d15e	0.004	0.003	−1.984	6.253	2.368	2.000
m5a8f01	0.006	0.003	−2.205	9.101	3.534	4.000
p5j2j	−0.004	−0.003	−2.103	5.773	2.161	2.000
hv3c1c	−0.003	−0.003	−0.264	6.572	2.880	3.000
p5q1a	0.004	0.003	−1.104	9.502	4.385	4.440
hv3j11	−0.005	−0.002	−5.206	9.047	2.335	2.145
hv4a24	0.004	0.002	−5.910	24.000	0.477	0.000
m4b4a2	0.004	0.002	−1.934	11.672	4.777	5.000
hv4g23j	−0.004	−0.002	−6.436	11.239	2.509	2.520
f3c3g	0.003	0.002	−2.749	11.868	4.951	5.000
cf2b_age	0.006	0.001	1.979	32.762	16.145	16.000
p5q1m	0.002	0.001	1.000	13.107	7.339	8.000
p5q3bw	−0.001	−0.001	−0.340	3.411	1.468	1.050
f5g0	−0.001	−0.001	−0.790	4.413	1.823	2.000
m3k22	−0.001	−0.001	−2.484	30.000	1.568	1.000
t5b4m	0.001	0.001	−2.168	4.762	1.023	1.000
p5q1j	0.001	0.001	−2.471	13.259	5.003	5.000
t5c13a	0.000	0.000	−0.705	6.486	2.840	3.000
m5g2c	0.000	0.000	0.490	3.156	1.780	2.000
p5q1n	0.000	0.000	−0.466	15.045	6.985	8.000
p5j11	0.001	0.000	−29.742	101.000	1.670	1.000
p5i34	0.000	0.000	−11.099	30.000	2.412	2.000
hv4pvceil	0.000	0.000	1.000	13.402	7.137	7.000
f1j13b	0.001	0.000	−74.150	114.000	8.340	2.000
hv3h2b	0.000	0.000	−9.969	32.327	9.015	9.000
hv5_wj9pr	−0.001	0.000	−42.993	134.028	37.099	35.000
f4l5d	0.000	0.000	−32.979	102.000	4.296	2.127
hv4mhtcm	0.000	0.000	58.054	188.954	161.872	161.898
f3k22	0.000	0.000	0.177	83.052	44.671	52.000
hv4wjpr22	−0.001	0.000	−63.172	139.146	49.741	52.000
m2d3b7	0.000	0.000	−71.132	103.000	8.064	2.000
hv5_ppvtpr	0.000	0.000	−49.223	137.714	36.153	32.000
hv5_wj10pr	0.000	0.000	−34.456	156.657	47.855	47.000
p5j10	0.000	0.000	−125.486	263.493	59.553	46.494
hv4k2_expen	0.000	0.000	−615.220	3,200.000	309.194	300.000
cm1hhinc	0.000	0.000	0.000	150,102.930	32,975.982	23,911.452
m3l1	0.000	0.000	−68,626.380	164,744.962	34,456.792	29,402.380

Note: The first column shows the variable names as in the original data set and codebooks. The second and third columns present regression coefficients from the least absolute shrinkage and selection operator. Coefficients in the second column are in original scale, while those in the third column are standardized. Columns 4 to 7 show the summary statistics for each variable.

Table A4. Summary of Variables Selected out of the Prediction Model for Layoff.

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
cm3alc_case	2.150	53.297	−0.105	1.000	0.001	0.000
cm4fdiff	2.097	42.186	−0.136	1.000	0.003	0.000
t5e7_3	−0.640	−22.264	1.000	2.077	2.000	2.000
m3d0	0.967	9.612	0.721	2.000	1.011	1.000
cm3span	0.738	3.054	−0.556	1.000	0.065	0.000
hv3a27d	0.232	1.314	0.000	1.571	0.967	1.000
m5f8a3	−0.275	−1.026	1.000	2.769	1.924	2.000
m5f7a	−0.274	−0.987	1.000	2.758	1.919	2.000
hv4r10a_2	−0.073	−0.899	−0.271	1.000	0.006	0.000
t5a9p	−0.169	−0.839	1.000	2.628	1.959	2.000
m3i25	0.225	0.810	1.000	2.896	1.914	2.000
f4i23m	−0.338	−0.692	−0.230	2.991	1.598	1.825
t5a4	0.191	0.674	0.187	2.312	1.093	1.000
hv3b7_3	0.242	0.649	−1.065	1.326	0.164	0.000
m3i23d	−0.266	−0.637	0.705	3.056	1.777	2.000
m3i8a3	−0.150	−0.611	1.000	2.944	1.934	2.000
o5a6a	−0.133	−0.412	0.923	2.941	1.876	2.000
hv4d2	−0.162	−0.374	−1.210	1.509	0.243	0.000
m4i7f	−0.129	−0.296	0.281	3.236	1.751	2.000
f4c7e	0.135	0.294	−0.172	3.076	1.370	1.224
m3i0q	0.165	0.260	−0.639	3.327	1.540	1.183
m3k27a	0.054	0.257	1.000	2.681	1.953	2.000
m5e8_7	−0.065	−0.246	−0.757	1.002	0.076	0.000
f4j4	0.114	0.237	−0.248	3.341	1.421	1.300
f4b5	0.104	0.222	0.253	3.278	1.676	1.845
f3i23e	0.088	0.219	0.468	3.041	1.801	2.000
hv3s4	0.052	0.206	0.002	3.000	1.044	1.000
m2g5	−0.100	−0.203	0.029	3.082	1.589	2.000
t5e15b	0.103	0.198	−0.573	4.000	1.249	1.000
hv4c1a	0.153	0.198	0.591	6.011	3.470	3.986
m3b4c	−0.103	−0.193	0.000	8.587	6.915	7.000
m5f23k	−0.076	−0.185	0.546	2.990	1.794	2.000
m3j0a	0.134	0.182	−0.659	4.053	1.712	2.000
t5c5	0.038	0.170	0.261	2.000	1.057	1.000
f3k14b	0.073	0.154	0.095	3.188	1.643	2.000
m4b4b1	0.090	0.142	−1.368	2.430	0.493	0.000
m2h18	−0.058	−0.136	0.290	3.080	1.714	2.000
m5i4	0.063	0.131	−0.138	2.780	1.388	1.000
m3i7f	−0.031	−0.117	1.000	2.682	1.923	2.000
p5h1	0.094	0.114	−1.302	5.000	1.646	1.313
p5j2e	0.098	0.102	−2.222	5.000	0.549	0.000
hv3e0b	−0.038	−0.077	−1.070	2.019	0.550	0.616
cm4marp	0.014	0.070	−0.555	1.000	0.039	0.000
m5e3	0.020	0.059	0.126	2.344	1.137	1.000
p5i18b	0.050	0.055	−1.737	6.000	0.866	1.000
f3k14e	0.023	0.048	0.133	3.080	1.662	2.000
m4j0	0.034	0.048	−0.248	4.036	1.685	2.000
f4h1q	0.029	0.043	−1.562	5.000	1.373	1.000
m4b4b19	0.017	0.042	−1.001	2.000	0.142	0.000
k5g2d	−0.040	−0.039	−2.837	4.287	0.911	1.000
p5l17d	0.015	0.035	0.520	3.055	1.777	2.000
m2j1	−0.036	−0.034	−0.318	5.528	2.223	2.000
hv4b9	0.014	0.030	−0.680	2.290	0.628	0.943

(continued)

Table A4. (continued)

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
p5l15	0.011	0.030	0.722	3.039	1.843	2.000
hv3j11	0.051	0.024	−5.206	9.047	2.335	2.145
m5e6	0.011	0.023	−0.005	2.850	1.417	1.141
m3i6a	−0.009	−0.021	0.602	3.249	1.791	2.000
p5q2d	−0.024	−0.021	1.000	12.038	7.776	8.000
p5i3	0.032	0.020	−3.120	10.000	2.318	2.000
p5q1m	−0.028	−0.015	1.000	13.107	7.339	8.000
cmf5fevjail	−0.008	−0.015	−1.051	2.149	0.481	0.414
m4b4a2	−0.030	−0.014	−1.934	11.672	4.777	5.000
f3c3g	−0.028	−0.013	−2.749	11.868	4.951	5.000
m5b30	−0.005	−0.011	0.285	3.494	1.725	2.000
hv3k3f	0.011	0.010	−1.513	5.367	1.789	1.433
k5b2b	−0.011	−0.009	−2.688	5.044	1.007	1.000
m3i18	−0.002	−0.005	0.601	2.945	1.803	2.000
p5h16a	0.006	0.003	−4.673	15.000	1.650	1.000
m5i3b	−0.001	−0.003	0.144	3.598	1.717	2.000
hv4mhtcm	−0.011	−0.001	58.054	188.954	161.872	161.898
f3k22	−0.004	0.000	0.177	83.052	44.671	52.000
f5k7	0.001	0.000	−10.113	30.000	2.816	2.000
f2g1a	0.001	0.000	−107.895	180.060	19.950	1.000
p5j10	−0.001	0.000	−125.486	263.493	59.553	46.494
m5j1	0.000	0.000	−67,800.515	175,267.257	42,014.790	36,344.900

Note: The first column shows the variable names as in the original data set and codebooks. The second and third columns present regression coefficients from the least absolute shrinkage and selection operator. Coefficients in the second column are in original scale, while those in the third column are standardized. Columns 4 to 7 show the summary statistics for each variable.

Table A5. Summary of Variables Selected out of the Prediction Model for Job Training.

	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
t5e7_3	−0.019	−0.674	1.000	2.077	2.000	2.000
m5e9_0	0.029	0.219	−0.438	1.000	0.018	0.000
m3i23c	−0.021	−0.149	1.000	2.469	1.979	2.000
m3i7f	−0.027	−0.102	1.000	2.682	1.923	2.000
m4i23d	−0.031	−0.091	0.640	3.130	1.865	2.000
m5f23a	−0.028	−0.086	0.850	2.909	1.880	2.000
p5l12b	−0.014	−0.076	1.000	2.697	1.966	2.000
m3k27a	−0.013	−0.060	1.000	2.681	1.953	2.000
m5f23e	−0.025	−0.055	0.309	3.271	1.700	2.000
hv3p6_e	0.012	0.055	−0.684	1.000	0.049	0.000
m4k26a	−0.013	−0.051	1.000	2.710	1.933	2.000
f3i6a	−0.017	−0.050	0.629	2.892	1.853	2.000
m5f23c	−0.019	−0.049	0.427	2.875	1.805	2.000
m4i23n	−0.019	−0.049	0.805	3.063	1.813	2.000
m5i14a3	−0.014	−0.044	0.889	2.907	1.881	2.000
m2h18	−0.017	−0.041	0.290	3.080	1.714	2.000
m5f23k	−0.016	−0.039	0.546	2.990	1.794	2.000
f5g28	−0.012	−0.037	0.837	2.906	1.882	2.000
m3i23d	−0.014	−0.033	0.705	3.056	1.777	2.000
m3b24	−0.011	−0.032	0.723	2.960	1.870	2.000
hv4l59	0.010	0.031	−0.987	2.000	0.102	0.000
f5i14b4	0.015	0.031	0.099	3.506	1.636	1.841
m5f23g	−0.014	−0.030	0.169	3.323	1.703	2.000
hv3s1_1	0.005	0.028	−0.485	1.000	0.026	0.000

(continued)

Table A5. (continued)

	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
hv3d2	0.013	0.027	−1.123	1.651	0.306	0.000
p5q3ag	0.008	0.026	0.119	3.000	1.081	1.000
m3l6a	0.004	0.024	0.301	2.000	1.056	1.000
m2h9a1	−0.010	−0.023	0.520	3.057	1.751	2.000
p5q3k	0.010	0.023	−0.408	3.000	1.207	1.000
p5q3bl	0.008	0.022	−0.103	3.000	1.128	1.000
p5h15	−0.009	−0.020	0.331	2.982	1.726	2.000
f2a7d	−0.009	−0.019	0.120	2.890	1.567	1.761
m2g5	−0.009	−0.018	0.029	3.082	1.589	2.000
hv4k9	0.007	0.017	−1.018	2.106	0.744	1.000
p5q3bo	0.007	0.017	−0.038	3.000	1.206	1.000
f2g13	−0.011	−0.016	−0.741	3.662	1.561	1.451
m5g0	0.011	0.015	−0.241	4.161	1.734	2.000
p5q3bn	0.007	0.015	−0.288	3.000	1.203	1.000
cm4marp	0.003	0.013	−0.555	1.000	0.039	0.000
m4c38	−0.003	−0.012	1.000	2.765	1.923	2.000
m3i23e	−0.005	−0.011	0.619	3.253	1.770	2.000
f4l6	0.004	0.010	0.015	2.634	1.207	1.000
m4b2	0.007	0.009	−0.761	5.000	1.529	1.000
m4i9	−0.003	−0.007	0.717	3.235	1.816	2.000
m2h19h	−0.002	−0.007	0.937	2.834	1.875	2.000
m3k3c	−0.001	−0.006	1.000	2.640	1.939	2.000
m5e6	0.002	0.004	−0.005	2.850	1.417	1.141
m5g16b	0.003	0.004	0.669	5.591	3.400	4.000
p5q3by	0.001	0.004	0.152	3.000	1.104	1.000
hv4f1f	−0.004	−0.004	−0.213	8.296	3.632	4.000
k5f1	0.002	0.003	8.032	12.466	9.995	9.814
f4b4b2	−0.002	−0.003	−1.793	3.315	0.699	0.759
hv3m2b	0.003	0.003	−1.787	3.509	0.818	1.000
cm5edu	0.003	0.003	−1.086	6.170	2.513	2.936
m5g2c	−0.001	−0.003	0.490	3.156	1.780	2.000
p5q2d	−0.003	−0.003	1.000	12.038	7.776	8.000
m5e8_5	−0.001	−0.003	−1.063	1.989	0.545	0.684
m4j0	0.001	0.002	−0.248	4.036	1.685	2.000
p5q3dk	−0.001	−0.002	0.061	4.241	2.282	2.000
m5g1	0.002	0.002	−0.966	5.401	2.427	2.137
p5j2j	0.002	0.002	−2.103	5.773	2.161	2.000
cf4povcab	−0.002	−0.001	−1.519	8.159	3.313	3.151
f5k3b	0.001	0.001	−1.222	4.284	1.233	1.171
p5q3dd	0.000	−0.001	0.312	4.576	2.434	2.592
k5e1d	−0.001	−0.001	−0.073	7.176	3.380	4.000
hv3j11	0.001	0.001	−5.206	9.047	2.335	2.145
hv3j7	0.001	0.000	−3.765	9.047	2.709	3.000
p5i31h	−0.001	0.000	−1.969	8.903	3.356	3.643
f3c3g	0.000	0.000	−2.749	11.868	4.951	5.000
p5j11	0.000	0.000	−29.742	101.000	1.670	1.000
p5q1j	0.000	0.000	−2.471	13.259	5.003	5.000
f3i4	0.000	0.000	−465.165	1,605.001	495.466	484.229

Note: The first column shows the variable names as in the original data set and codebooks. The second and third columns present regression coefficients from the least absolute shrinkage and selection operator. Coefficients in the second column are in original scale, while those in the third column are standardized. Columns 4 to 7 show the summary statistics for each variable.

Table A6. Summary of Variables Selected out of the Prediction Model for Eviction.

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
t5e7_3	−4.025	−139.953	1.000	2.077	2.000	2.000
hv3s1_3	5.256	77.284	−0.194	1.000	0.005	0.000
o5notinhouse	−1.212	−6.675	1.000	2.534	1.967	2.000
m5c1d	0.873	6.014	0.534	3.000	1.015	1.000
m5c1e	1.014	5.607	0.436	3.000	1.025	1.000
hv3a11	1.057	4.792	−0.664	1.000	0.050	0.000
m3l6a	0.740	4.013	0.301	2.000	1.056	1.000
m5i3c	−0.914	−3.380	1.000	2.899	1.924	2.000
hv3p6_e	0.702	3.261	−0.684	1.000	0.049	0.000
m5f23c	−1.095	−2.839	0.427	2.875	1.805	2.000
m3k3c	−0.593	−2.488	1.000	2.640	1.939	2.000
m5e8_7	0.640	2.405	−0.757	1.002	0.076	0.000
hv3t1	−0.553	−2.078	0.000	1.818	0.927	1.000
m4k26a	−0.475	−1.889	1.000	2.710	1.933	2.000
m5f23k	−0.754	−1.843	0.546	2.990	1.794	2.000
f5g23	−0.391	−1.765	1.000	2.776	1.944	2.000
t5c15	0.529	1.710	0.893	3.032	1.883	2.000
f3i6a	−0.542	−1.581	0.629	2.892	1.853	2.000
hv3s4	0.399	1.579	0.002	3.000	1.044	1.000
m4c38	−0.416	−1.551	1.000	2.765	1.923	2.000
hv3s1_1	0.263	1.516	−0.485	1.000	0.026	0.000
t5c5	−0.318	−1.423	0.261	2.000	1.057	1.000
f4i23d	−0.426	−1.318	0.586	2.812	1.877	2.000
t5a9p	−0.264	−1.311	1.000	2.628	1.959	2.000
p5q3bp	−0.231	−1.146	0.449	3.000	1.029	1.000
hv4l59	0.376	1.115	−0.987	2.000	0.102	0.000
m5f7b	−0.468	−1.105	0.134	3.352	1.767	2.000
f4l6	0.404	1.070	0.015	2.634	1.207	1.000
m1c1d	0.127	1.061	0.878	3.000	1.012	1.000
f5g3	−0.349	−0.906	0.436	3.038	1.817	2.000
m2h19h	−0.297	−0.889	0.937	2.834	1.875	2.000
f5a8	0.154	0.844	0.466	2.000	1.038	1.000
hv4l29	−0.275	−0.842	−0.957	2.000	0.103	0.000
m3i25	−0.224	−0.808	1.000	2.896	1.914	2.000
o5g7	0.390	0.785	0.092	3.232	1.555	1.727
m5f23b	−0.199	−0.785	1.000	2.658	1.933	2.000
m3i6a	−0.315	−0.777	0.602	3.249	1.791	2.000
m3l3	−0.342	−0.705	−0.698	2.808	1.371	1.000
m1g4	−0.331	−0.673	1.000	4.323	3.772	4.000
m2b9	−0.212	−0.585	0.703	2.730	1.842	2.000
p5q3bb8	0.086	0.573	0.565	3.000	1.016	1.000
p5q3af	−0.196	−0.550	−0.178	3.000	1.121	1.000
p5q3k	0.235	0.526	−0.408	3.000	1.207	1.000
m5f7a	0.144	0.520	1.000	2.758	1.919	2.000
p5q3by	0.161	0.485	0.152	3.000	1.104	1.000
cmf5fevjail	−0.225	−0.448	−1.051	2.149	0.481	0.414
m5g2c	−0.177	−0.423	0.490	3.156	1.780	2.000
p5l15	−0.146	−0.398	0.722	3.039	1.843	2.000
m4f2f1	0.167	0.397	−0.576	2.645	1.257	1.000
f4h1q	0.264	0.392	−1.562	5.000	1.373	1.000
t5d1a	0.196	0.386	−0.473	3.000	1.238	1.000
t5a9o	−0.077	−0.379	1.000	2.578	1.955	2.000
f2a7d	−0.181	−0.363	0.120	2.890	1.567	1.761
o5f4	0.155	0.288	1.000	6.427	4.796	5.000

(continued)

Table A6. (continued)

Variable	glmnet Coefficient	Rescaled Coefficient	Minimum	Maximum	Mean	Median
hv4a1	0.186	0.239	−0.959	5.000	1.612	1.416
f5i14b4	0.104	0.217	0.099	3.506	1.636	1.841
f4i0n2	0.182	0.211	−0.666	5.538	2.041	2.000
p5q3cg	−0.098	−0.195	−0.249	3.000	1.283	1.000
f4b4b2	−0.131	−0.187	−1.793	3.315	0.699	0.759
m4b4b1	−0.104	−0.165	−1.368	2.430	0.493	0.000
t5b1o	0.138	0.157	−0.299	5.953	2.703	2.821
f4b5	−0.067	−0.143	0.253	3.278	1.676	1.845
hv3c1c	0.142	0.139	−0.264	6.572	2.880	3.000
p5q3a	−0.069	−0.137	−0.310	3.006	1.322	1.000
f5g19	−0.127	−0.133	−2.271	4.391	1.176	1.000
f5k3b	0.091	0.117	−1.222	4.284	1.233	1.171
hv3m49	0.078	0.115	−1.510	2.724	0.696	0.970
hv4f1f	−0.136	−0.112	−0.213	8.296	3.632	4.000
f5k14a	0.114	0.093	−2.372	5.166	1.608	1.573
hv3m44	−0.052	−0.091	−1.514	2.409	0.484	0.331
p5m2e	0.090	0.091	−2.001	4.933	1.794	1.703
m5b22b	0.108	0.088	−1.816	6.664	2.647	2.730
m2j1	0.081	0.077	−0.318	5.528	2.223	2.000
m4r3	−0.028	−0.068	0.283	2.867	1.787	2.000
cm5edu	0.066	0.066	−1.086	6.170	2.513	2.936
p5h15	−0.028	−0.063	0.331	2.982	1.726	2.000
hv3j19	−0.021	−0.060	−1.233	7.000	0.023	0.000
p5q2d	−0.068	−0.059	1.000	12.038	7.776	8.000
o5f3	0.050	0.059	1.000	7.366	4.509	5.000
hv4d17	0.077	0.051	−2.354	6.721	2.136	1.924
hv4d15c	0.067	0.047	−2.936	7.712	2.628	2.632
m5g24	0.018	0.047	0.457	2.909	1.810	2.000
k5g1c	−0.041	−0.046	−0.838	5.103	2.159	2.000
p5q1a	0.067	0.043	−1.104	9.502	4.385	4.440
k5e2c	−0.032	−0.043	−2.183	4.000	0.264	0.000
f4j2	−0.013	−0.042	0.829	2.897	1.891	2.000
p5q1m	−0.059	−0.031	1.000	13.107	7.339	8.000
k5g2f	0.035	0.030	−2.164	4.713	1.237	1.000
k5a1b	0.028	0.027	−1.681	5.260	2.193	2.633
hv3j11	0.047	0.021	−5.206	9.047	2.335	2.145
p5k1e	0.015	0.021	−0.684	4.344	1.909	2.000
p5j7a	0.016	0.019	0.521	6.753	3.466	4.000
hv3c8	0.017	0.017	−1.376	5.377	1.815	1.990
p5q3bn	−0.007	−0.016	−0.288	3.000	1.203	1.000
cm2povco	0.024	0.015	−2.858	6.658	1.727	1.314
f5g0	0.010	0.013	−0.790	4.413	1.823	2.000
cm3hhimp	−0.020	−0.013	−1.784	7.424	2.588	2.549
k5a3c	0.013	0.012	−1.840	5.302	1.872	2.000
m3i23d	−0.003	−0.008	0.705	3.056	1.777	2.000
f3c3g	0.010	0.005	−2.749	11.868	4.951	5.000
m5g0	0.003	0.005	−0.241	4.161	1.734	2.000
p5i30a	−0.002	−0.004	0.059	3.247	1.633	2.000
m2d3b7	−0.003	0.000	−71.132	103.000	8.064	2.000
hv3whp	0.003	0.000	−34.769	162.316	63.259	65.140
p5j10	−0.002	0.000	−125.486	263.493	59.553	46.494
cm5hhinc	0.000	0.000	−72,839.080	165,385.252	41,747.160	34,263.732

Note: The first column shows the variable names as in the original data set and codebooks. The second and third columns present regression coefficients from the least absolute shrinkage and selection operator. Coefficients in the second column are in original scale, while those in the third column are standardized. Columns 4 to 7 show the summary statistics for each variable.

Table A7. Variables Selected as Predictive across Models.

Grit . . . GPA	Hardship . . . Eviction . . . Layoff . . . Training
m3i23d	m3k22
m5f23k	m3l3
p5q2d	m4f2e2
hv3j11	k5g2h
t5e7_3	m5g19
f3c3g	m5j2
	p5i14
	p5m1
	hv5_ppvtpr
	hv5_wj10pr
	t5b1d
	t5b1u
	t5b3e
	t5c16
	cf5povco
	f2k12
	m3b5
	m3i8a3
	cm4marp
	k5a3c
	k5b1b
	m5f23c
	f5k14b
	p5m2e
	p5q3bb8
	o5f4
	o5f6
	hv4d15c
	hv4l47
	hv4r10a_3
	m5e9_0
	m5i3c
	p5j10

Note: Column 1 shows the intersection of predictive variables between the final model for two outcomes: grit and GPA. Column 2 shows the intersection of predictive variables among the final model for four outcomes: hardship, eviction, layoff, and job training. No variables were selected across all six outcomes.

Table A8. Out-of-sample Results of Predictions for Six Outcomes.

Submissions	Test Data	Material Hardship	GPA	Grit	Eviction	Layoff	Job Training
First	Leaderboard	0.024	0.391	0.222	18.292	3.354	2.921
First	Holdout	0.019	0.358	0.256	17.305	3.430	2.803
Second	Leaderboard	0.025	0.398	0.224	0.056	0.180	0.197
Second	Holdout	0.019	0.359	0.256	0.058	0.171	0.181
Second seeded	Leaderboard	0.024	0.382	0.229	0.059	0.185	0.202
Second seeded	Holdout	0.019	0.361	0.253	0.059	0.167	0.181

Note: Reported numbers are mean squared errors (MSEs). The first two rows show the results from our initial submission, and the third and fourth rows show our award-winning results (second submission). The two rows under "second seeded" are results that are ready for replication. "Leaderboard" refers to a temporary validation data set, the MSE from which was immediately available to participants. "Holdout" refers to another testing data set upon which final performance among participants was evaluated. GPA = grade point average.

Table A9. Union over the Six Sets of Remaining Variables after Preprocessing Stage.

m1i1
m1i3
f1b20
cm2povco
f2g1a
f2h5a
m3k22
m3l3
f3b3
f3k12
m4f2e2
m4f2f1
k5g2d
k5g2h
k5g2m
m5b3
m5b30
m5g19
m5j1
m5j2
f5i13
f5k7
p5i14
p5i23
p5i30a
p5i31h
p5i34
p5j4b
p5l13f
p5m1
p5q3u
p5q3bt
p5q3bw
p5q3cg
hv5_ppvtpr
hv5_wj10pr
t5b1d
t5b1f
t5b1u
t5b1w
t5b3e
t5c16
cm5fevjail
cmf5fevjail
cm5edu
cf5povco
hv3c5
hv3c8
hv3g1f
hv3m2b
hv3m2c
hv3m7
hv3m21
cm1bsex
m1b12d
m1c1d
cm1hhinc
f1j13b
m2d2
m2d2c
m2d3b5
m2d3b7
m2f5
cf2b_age
f2k12
m3b5
m3i8a3
m3i18
m3i23c
m3l1
f3i6h
f3i23e
f3k14e
m4b4b1
m4b6c
cm4marp
m4h3a
f4b4b2
f4b5
f4c7e
f4i0n1
f4i0n5
k5a1b
k5a2f
k5a3c
k5b1b
k5b2b
k5d1h
k5e1c
k5e2c
k5g1b
k5g1c
k5g1e
k5g2f
m5a5b01
m5a8f01
m5b22b
m5e1k
m5f23c
m5g16b
m5g24
m5g28
m5g31
m5i16c
m5e8_5
m5e8_7
f5a7
f5c1f
f5g0
f5g3
f5g23
f5g28
f5k3b
f5k14b
f5e9_4
p5h15
p5i1f
p5i1j
p5i26
p5i33b
p5j1
p5j2e
p5j2j
p5j7a
p5k1e
p5l15
p5l17d
p5m2c
p5m2e
p5q1j
p5q1n
p5q3a
p5q3d
p5q3af
p5q3bb8
p5q3bp
p5q3cb
hv5_wj9pr
o5a2
o5f3
o5f4
o5f6
t5a9p
t5b1o
t5b4m
t5c15
t5e11
t5e15b
cm5md_case_lib
hv3a27d
hv3e0b
hv3h2b
hv3j19
hv3k3f
hv3m44
hv3r12
hv3v6b
hv3whp
hv4a24
hv4b9
hv4d1a
hv4d2
hv4d15c
hv4d15e
hv4d17
hv4g23j
hv4l42
hv4l47
hv4r10a_2
hv4r10a_3
hv4sex_child
hv4k2_expen
hv4pvceil
hv4pverr
hv4wjpr22
m2g5
m2h9a1
m2h18
f2g13
m3b24
m3i7f
m3i23d
m3i23e
m3k27a
f3i6a
m4b2
m4i7f
m4i9
m4i15
m4i23d
m4i23h
m4i23n
m4j0
m4k26a
cm4hhinc
cm4povco
f4l6
cf4povcab
k5e1d
m5e6
m5f23a
m5f23e
m5f23g
m5f23k
m5g0
m5g1
m5i14a3
m5j6h
m5e9_0
f5i14b4
p5h13
p5h14
p5j11
p5l12b
p5q2d
p5q3k
p5q3ag
p5q3bl
p5q3bn
p5q3bo
p5q3by
hv3d2
hv3j7
hv3j11
hv4f1f
hv4k9
hv4l59
k5f1
m1g4
m2b9
m2h19h
m2j1
m3i6a
m3i25
m3k3c
m3l6a
cm3hhimp
f3i4
m4c38
f4h1q
f4i0n2
f4i23d
f4j2
m5c1d
m5c1e
m5f7b
m5f23b
m5g2c
m5i3c
f5a8
f5g19
p5q1a
o5g7
o5notinhouse
t5a4
t5d1a
t5e7_3
hv3a11
hv3c1c
hv3m49
hv3p6_e
hv3r5
hv3s1_1
hv3s1_3
hv3s4
hv3t1
hv4a1
m3b4c
m3d0
m3i0q
m3j0a
cm3alc_case
cm3span
f3c3g
f3k14b
f3k22
cm4fdiff
m4b4a2
m4b4b19
f4i23m
f4j4
m5e3
m5f7a
m5f8a3
f5k2a
f5k14a
p5h1
p5i3
p5i18b
p5j10
p5q1m
o5a6a
t5a9o
t5c5
t5c13a
cm5span
cm5hhinc
cf5hhinc
hv3b7_3
hv4c1a
hv4mhtcm
m1i2b
m2g8
f2a7d
m3a13
m3i23h
m3k3b
m3l2
f3d1
m4b8d
m4r1
m4r3
m4k3b
m4l2
cf4cohp
f4k3b
f4l5d
m5h3
m5i1
m5i3b
m5i4
m5i13
f5g16c
f5i13p
f5e9_7
p5h16a
p5i20c
p5j2g
p5l8_6
p5q3at
p5q3bb5
p5q3dd
p5q3dh
p5q3dk
o5d1_2
o5d1_6
t5d8a
hv4l13
hv4l29

show the results from our initial submission, and the third and fourth rows show our results that achieved the lowest MSE for material hardship (second submission). The two rows under "second seeded" are results that are ready for replication. "Leaderboard" refers to a temporary validation data set. Its MSE was immediately available to participants. "Holdout" refers to another testing data set upon which final performance across participants is evaluated.

The difference of performances (MSE) between the first submission and the second submission is in binary outcomes. For the first submission, we used log odds as a prediction outcome, but the FFC required us to submit probabilities as an outcome. The second submission corrected this step. The first two submissions, "first" and "second," were not properly seeded. The MSE for material hardship from the holdout data in the "second" submission was the lowest among all predictions in the FFC and is the one discussed in the main text. The "second seeded" submission used the same exact code from "second" but was properly seeded and thus fully reproducible. The holdout MSE for material hardship in the seeded submission is identical to the one in the unseeded (lowest in the FFC) up to three decimal places.

Authors' Note

The authors contributed equally to this article. The previous version of this article was discussed in the FFC Scientific Workshop.

Acknowledgments

We thank the two editors of the special collection and three anonymous reviewers for detailed comments and feedback. We are grateful to the Fragile Families and Child Wellbeing Study for sharing with us the data for this study.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for the Fragile Families and Child Wellbeing Study was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development through grants R01HD36916, R01HD39135, and R01HD40421 and by a consortium of private foundations, including the Robert Wood Johnson Foundation. Funding for the FFC was provided by the Russell Sage Foundation. We use R version 3.4.3 (R Core Team, 2017) and the following R packages for the analysis: glmnet version 2.0.13 (Simon et al. 2011) and Amelia version 1.7.4 (Honaker et al. 2011).

Supplemental Material

Supplemental material for this article is available with the manuscript on the *Socius* website.

References

Greenhill, Brian, Michael D. Ward, and Audrey Sacks. 2011. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models." *American Journal of Political Science* 55(4): 991–1002.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand, and Henrik Urdal. 2013. "Predicting Armed Conflict, 2010–2050." *International Studies Quarterly* 57(2): 250–70.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(7):1–47.

King, Gary, James Honaker, Anne Joseph O'Connell, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69.

Mayer, Susan E., and Christopher Jencks. 1989. "Poverty and the Distribution of Material Hardship." *Journal of Human Resources* 24(1):88–114.

R Core Team. 2017. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing.

Reichman, Nancy E., Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. 2001. "Fragile Families: Sample and Design." *Children and Youth Services Review* 23(4–5):303–26.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, and Sara McLanahan. 2019. "Introduction to the Special Collection on the Fragile Families Challenge." *Socius* 5. doi:10.1177/2378023119871580.

Simon, Noah, Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39(5):1–13.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society, Series B (Methodological)* 58(1):267–88.

Author Biographies

Diana Stanescu is a PhD candidate in the Department of Politics at Princeton University and an exchange scholar in the Department of Government at Harvard University. At Princeton, she holds a Center for International Security Studies and a Bradley Research Program Fellowship and is a student associate of the Niehaus Center for Globalization and Governance. She studies international trade, regulation, and lobbying with a focus on Japan using formal and quantitative methods. Her research addresses the overarching question of how the structure of foreign policy decision making shapes international cooperation. Previously she worked in international development, designing methodologies to evaluate programs on public policy, labor market discrimination, and gender equity.

Erik Wang is a PhD candidate in the Department of Politics at Princeton University. He holds a Quantitative and Analytical Political Science Fellowship at Princeton. He studies comparative political economy with a particular interest in state building, bureaucracy, and corruption. His research addresses the spatial-temporal variation in state capacity across local governments in China. He also does research on statistical methods of causal inference, with a focus on time-series cross-sectional data.

Soichiro Yamauchi is a graduate student in the Department of Government at Harvard University. He is broadly interested in political methodology and computational social science. In particular, he focuses on the large-scale record linkage problem and causal inference in panel data.