Estimating Public Assistance Program Participation*

Elias Ilin† Alvaro Sanchez‡ Victor Yifan Ye§

Abstract

Many low-income families are eligible for public assistance but do not "take up" the benefit. In this paper, we develop a novel, machine learning-based method to estimate take-up rates for seven major U.S. social safety net programs. Our method consists of four steps. First, we determine the universe of respondents to the 2019 Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) that are eligible for social safety net programs of interest. Second, we contrast three alternative algorithms for binary classification (gradient-boosted trees, random forests, and logistic regression) and evaluate their performance on predicting out-of-sample program take-up in the ASEC. Specifically, we select XGboost (Chen and Guestrin 2016), a gradient-boosted classifier, for its relative efficiency and high performance of 70 to 90 percent out-of-sample accuracy. Third, to address underreporting of program participation, for each program-eligible observation in ASEC that does not report participation, we utilize XGboost to assign participation indicators to match national statistics on take-up rates. Finally, we apply a similar algorithm to the adjusted ASEC data, and re-estimate the model to relate social-economic factors to public assistance take-up. Importantly, the imputation method developed in this paper can be applied to any dataset with a detailed family-level demographic and income information. We demonstrate how the method can be used, by applying it to the 2019 American Community Survey (ACS) and describing patterns of the U.S. social safety net program participation. We find that take-up rates vary considerably by program, geographical location, and demographic characteristics.

*Paper prepared for the 115th Annual Conference on Taxation. Results are preliminary. Please contact the authors before citing.

†Federal Reserve Bank of Atlanta. elias.ilin@atl.frb.org

‡Federal Reserve Bank of Atlanta. alvaro.sanchez@atl.frb.org

§Opendoor Technologies and Stanford DEL. victorye@stanford.edu

1 Introduction

The main goals of social safety net programs, including providing those in need with a basic standard of living, can be seriously undermined if families do not take up the available support. This issue of low benefit take-up has received increasing attention from economists and public policy researchers (Anderson and Meyer 1997; Yaniv 1997; Riphahn 2001). The fact that families do not collect benefits available to them seems inconsistent with predictions based on economic theory. However, there is firm evidence that take-up rates of most of the U.S. social safety net programs are very low (Chien 2015; Kosar and Moffitt 2017; Foster and Rojas 2018; Giannarelli 2019).

For public policy to address the challenges created by low take up rates, it is first necessary to quantify the extent to which the issue exists and identify which programs are affected. Additionally, it requires an understanding of the determinants of low take-up rates. Most studies conducted so far rely on survey data to analyze low benefits take up, but these data sources include misreporting and measurement errors that affect the accuracy of the variables needed to estimate both benefits receipt and eligibility.

Data from nationally representative household surveys are widely considered unreliable for measuring the uptake of public assistance benefits (Coder and Scoon-Rogers 1996; Roemer 2000; Wheaton 2016; Rothbaum 2015). These data often bias take-up rates downward due to underreporting of receipts of benefits, either as a result of unit and item non-response, or inaccurate household responses. Ultimately, using these unreliable data has led to an underestimation of the true effect of public assistance on the income distribution (Lynch et al. 2008; Jolliffe et al. 2005) and an overestimation of poverty among low- and moderate-income households (Meyer and Wu 2018). Various methodological approaches have attempted to account for the underreporting of benefits take-up and receipt. However, the use of machine learning algorithms to correct for underreporting is a relatively new and novel approach that we attempt in this paper. The machine-learning approach contrasts with the majority of microsimulation literature that includes the use of linked administrative data.

Recent research has attempted to uncover the extent to which under-reporting is prevalent in household surveys and address the unreliability of the data through several methods. Meyer et al. (2009) compared household survey data with administrative data from nine public assistance programs, finding a rapid rise in under-reporting of transfer program receipts and dollars received over time. The study also found that in several prominent surveys (American Community Survey, Consumer Expenditure Survey, Current Population Survey, Panel Study of Income Dynamics) nearly half of Temporary Assistance for Needy Families and food stamp dollars went unaccounted for because of discrepancies between administrative and survey data.

Meyer et al. (2015) also analyzed five national household surveys and found an increasing share of underreporting and imputation in ten transfer programs through linked survey and administrative data. The analysis revealed the Consumer Expenditure Survey had the lowest reporting rate of benefits receipts, while Survey of Income and Program Participation showed the highest rate across programs. The authors theorized that unit and item non-response in each survey were driven by factors such as continuity of benefits receipt, ease of reporting, survey structure, and a reduction of interview lengths.

Survey data have been shown to impact the measurement of the income distribution and public assistance receipt for specific sub-populations. Bee and Mitchell (2017) used linked administrative data and found that the median household income of householders aged 65 and up was 30 percent higher than originally reported in CPS ASEC survey data. The same study found that the adjusted poverty rate of individuals ages 65 and up was 6.9 percent, according to linked administrative data, rather than 9.1 percent that was reported from survey data. The authors attributed the data discrepancies to a steadily increasing amount of retirement income under-reporting in household surveys over time.

Additional evidence from Meyer and Mittag (2019) suggests that certain public assistance programs (SNAP, TANF, New York General Assistance) have double the poverty-reducing effects than what can be estimated using CPS ASEC survey data alone, without linked

administrative data. The authors find that correcting for underreporting reduces poverty rates for members of disadvantaged groups, especially single mother-headed households by as much as 11 percentage points. Meyer et al. (2022) also analyzed the misreporting of Food Stamp Program receipts and find substantial underreporting across multiple household surveys. Their results emphasize that underreporting of participation is prevalent among single parents, non-White individuals, and the elderly population.

In this paper we develop and implement a novel machine-learning method to estimate the probability of programs take-up among eligible people for seven major U.S. social safety net programs: Supplemental Nutrition Assistance Program (SNAP), Section 8 Housing Choice Voucher Program, Medicaid for Adults, Medicaid for Children/Children Health Insurance Program (CHIP), Affordable Care Act (ACA) marketplace health insurance subsidies, Earned Income Tax Credit (EITC), and Child Tax Credit (CTC). Our machine-learning approach does not rely on linked administrative data, but rather uses survey data itself to determine the predominating demographic factors associated with the take up of public assistance benefits among eligible families, and subsequently estimate the probability of program participation given those factors.

We evaluate and contrast four alternative binary classification algorithms to estimate the probability of participation (gradient-boosted trees, random forests, and logistic regression with and without penalization). We favor XGboost (Chen and Guestrin 2016), a gradient-boosted classifier, for its relative efficiency and high performance of 70 to 90 percent out-of-sample accuracy. To account for the pervasive underreporting of program receipt, each program-eligible, but non-reporting CPS ASEC observation is assigned a participation indicator that aligns with national statistics on take-up for a given program. After adjusting for underreporting of program receipt, our model is then re-estimated with the adjusted ASEC data to get final estimates.

Our method to impute program participation has important practical applications for economics and public policy researchers. Model estimates can be easily applied to any

individual- or family-level dataset with detailed demographic and income information. We demonstrate how the model works by applying it to the 2019 American Community Survey (ACS). We establish five stylized facts about participation in the U.S. social safety net programs.

First, among all U.S. social safety net programs, the Child Tax Credit (CTC) has by far the largest number of eligible and participating families. Housing Choice Voucher Program has the second largest number of eligible families, but the lowest uptake. Second, there is a strong correlation between participation in the ACA marketplace health insurance subsidies and in the Housing Choice Voucher program. There is no statistically meaningful correlation between participation in Earned Income Tax Credit (EITC) and Medicaid for Children/CHIP. Third, Child Tax Credit as a single benefit is the most common benefit bundle. Combination of SNAP, Housing Choice Voucher, Medicaid for Adults, Medicaid for Children, ACA marketplace health insurance subsidies, EITC, and CTC is the second most common bundle, although only 4.3 percent of all families are estimated to receive it. Fourth, there is no clear relationship between demographic characteristics and program take up that holds across all programs. We find that this relationship is very program specific. Fifth, program eligibility, participation, and take-up rates vary significantly by state. For example, take-up rates of Supplemental Nutrition Assistance Program (SNAP) vary from 60 percent in California to 87 percent in South Dakota; SNAP eligibility rates vary from 14 percent in South Dakota to 33 percent in Kentucky.

Our novel machine learning-based technique developed in this paper, combined with the established stylized facts about social programs take-up, can inform future research which is thought to better understand the determinants of social safety net program participation.

2 Methodology

Our machine learning-based method to estimate the probability of programs take-up among eligible people consists of the following four steps.

1. **Step 1:** We determine the universe of respondents to the 2019 ASEC that are eligible for public assistance programs and tax credits of interest. Additionally, for each eligible family, we estimate the expected dollar value of the benefit.

2. **Step 2:** We estimate the model of take-up probability for each public assistance program of interest using the data on income, demographics, expected dollar value of the benefit, and *unadjusted* programs take-up.

3. **Step 3:** We use the estimates from Step 2 to assign the probability of take-up to each eligible unit CPS ASEC that does not report take-up and inflate the take-up numbers to match national averages.

4. **Step 4:** Finally, we repeat Step 2 and re-estimate the take-up probability model using *adjusted* program take-up data.

In the rest of this section we discuss each step in details.

2.1 Estimating Public Assistance Programs Eligibility for ASEC Respondents

As a first step, for each household in the 2019 ASEC we determine eligibility status and expected value of each of the following seven public assistance programs and tax credit: Supplemental Nutrition Assistance Program (SNAP), Section 8 Housing Choice Voucher, Medicaid for Adults, Medicaid for Children/Children Health Insurance Program (CHIP), Affordable Care Act (ACA) marketplace health insurance subsidies, Earned Income Tax Credit (EITC), and Child Tax Credit (CTC).

ASEC does not contain information on the family's participation in other important social safety net programs such as Child Care and Development Fund childcare subsidy program, Head Start, Low Income Home Energy Assistance Program, and others. However, seven programs that ASEC provides information for and this paper analyzes are by far the largest social safety net programs in terms of the share of total funds allocated to them. [1]

To determine eligibility status and implied dollar value of each benefit we use state-specific tax rules and transfer-payment eligibility rules as collected in the Policy Rules Database (Ilin and Terry 2021). The Policy Rules Database contains rules and provisions for all of the major federal and state assistance programs, tax rules, and tax credits available to working adults and their dependents. See Appendix A for additional details on the Policy Rules Database.

Table 1 reports participation rates (share of all families that report participate in a given program), eligibility rates (share of families that are program eligible), and take-up rates (share of eligible families that report program participation) for all seven programs.

Medicaid for Children has the highest income eligibility thresholds among all public assistance programs and the highest eligibility rate – 61.4 percent of all U.S. families are estimated to be eligible for this program. Medicaid for Adults, on the contrary, has one of the most strict eligibility requirement among social safety net programs and therefore the lowest share of eligible families (12.95 percent). Medicaid for Children also has the highest reported overall participation rate (34.2 percent) followed by CTC (20.9 percent). These two programs also have the highest take-up rates – 55.7 percent and 77.8 percent respectively. Housing Choice Voucher and ACA subsidy have high share of program eligible families (52.5 and 40.2 percent respectively), but very small share of eligible people actually participate in the program. The reported take-up rates for Housing Choice Voucher and ACA subsidy are only 7.8 and 3.5 percent respectively.

[1] See Center on Budget and Policy Priorities (2022)

Table 1: Estimated Family-Level Program Participation, Eligibility, and Take-up Among Respondents to the 2019 ASEC

	Participation Rate (%)	Share of Eligible (%)	Take-Up Rate (%)
SNAP	6.4	20.6	31.3
Housing Choice Voucher	4.2	52.5	7.8
Medicaid for Adults	5.7	13.0	44.3
Medicaid for Children	34.2	61.4	55.7
ACA Subsidy	1.4	40.2	3.5
EITC	8.4	18.3	45.7
CTC	20.9	26.9	77.8

Source: 2019 Annual Social and Economic Supplement (ASEC) and Policy Rules Database (Ilin and Terry 2021), authors' calculations
Note: Participation rate is a share of all families that are eligible for a program
Take-up rate is a share of eligible families that participate in a program

2.2 Model of Take-Up Probability Estimation

We use reported data on program participation from 2019 ASEC to estimate the model of take-up probability for each public assistance program of interest. We contrast four alternative algorithms for binary classification (gradient-boosted trees, random forests, and logistic regression) and evaluate their performance on predicting out-of-sample program take-up in the ASEC. Specifically, we select XGboost (Chen and Guestrin 2016), a gradient-boosted classifier, for its relative efficiency and balanced out-of-sample performance.

2.2.1 Alternate Models of Estimating Participation

We contrast four approaches to modeling take-up probabilities: two regression-based and two decision-tree-based. A unique model is tuned and trained for each type of model and welfare program. For all four approaches, our covariates includes age, sex of head of household, race, number of kids and number of family members, household income, marital status, education attainment, state, employment status, health insurance status, housing status, and an indicator for residency in a metro area.[2] We additionally include the respective value

[2]Collinear variables (e.g. value of ACA for imputing medicaid adult participation) are omitted. Factor inputs are transformed to levels that are represented in the ACS and ASEC, and top-coded to ensure a

of the seven studied welfare programs, as estimated via the PRD.[3] To prevent the "fuzzy" classification problem associated with continuous variables and decision trees, we transform all dollar-value variables (values of benefit programs, age, income) into discrete factors for decision-tree-based methods.[4] For regressions, we log-transform program benefit values and income, and include a quadratic term for age of head of household.

For each model, 20 percent of the data is excluded from training as a holdout set to evaluate out-of-sample performance. Our first approach is a simple logistic regression of reported participation on the aforementioned input variables. Our second approach utilizes an elastic net regression (Zou and Hastie 2005) with a binomial link. This is equivalent to a logistic regression but with two separate regularization terms: λ, which governs the overall shrinkage of regression parameters, and α, which governs the relative strength of L1 (LASSO) versus L2 (ridge) regularization.[5] We train the elastic net with 20-fold cross-validation over a fine grid of α and λ values.

Our third model, and the approach we choose for subsequent estimation procedures, utilizes XGBoost (Chen and Guestrin 2016), a boosted-tree-based classification algorithm that incorporates a number of optimization techniques and improvements compared to traditional decision tree methods. For our fourth approach, we train a traditional random forest (Ho 1995) model to provide a point of contrast. For both tree-based models, we perform hyperparameter tuning by grid search. Details of hyperparameter tuning are discussed in Appendix B.

consistent number of bins between the three datasets.

[3]For participants, these values should be the actual amount received from the program in question, and for non-participants they represent the value that they would have received, assuming participation.

[4]See Kotsiantis (2013) for an overview. When classifying on dollar-value (or age) predictors, tree-based methods will treat each dollar value (or year) as a unique bin, leading to unstable estimates. We use bins of \$10,000 for income and ACA value, bins of \$5,000 for the value of section 8, adult medicaid, and child medicaid, and bins of \$2,000 for SNAP, EITC, and CTC.

[5]See Tibshirani (1996) for a comparison between L1 and L2 regularization.

2.2.2 Results of the Take-up Models

Table 2 contrasts the out-of-sample performance of the four models for predicting SNAP participation. Performance contrasts for other six welfare programs are presented in Appendix B. Although the four models performance similarly in terms of accuracy of predicting SNAP participation, true positive and true negative rates differ significantly. The logistic regression performs worst in this regard, with a high sensitivity rate of 93.2 percent but low specificity of 29.4 percent, suggesting under-performance for predicting non-participants. XGBoost produces the highest specificity at 48.0 percent while also maintaining sensitivity in line with the random forest and elastic net approaches. Overall, the three machine learning-based models are relatively similar in terms of accuracy and sensitivity, with random forests being slightly less balanced in terms of performance. However, XGBoost is by far the most efficient, requring roughly a fifth of the training time of the elastic net model.[6]

Table 2: Comparison of OOS Performance of Alternate Imputation Models, SNAP

	Logit	Elastic Net	XGBoost	Random Forest
Accuracy	0.705	0.720	0.725	0.728
Sensitivity	0.932	0.858	0.855	0.882
Specificity	0.294	0.472	0.480	0.430

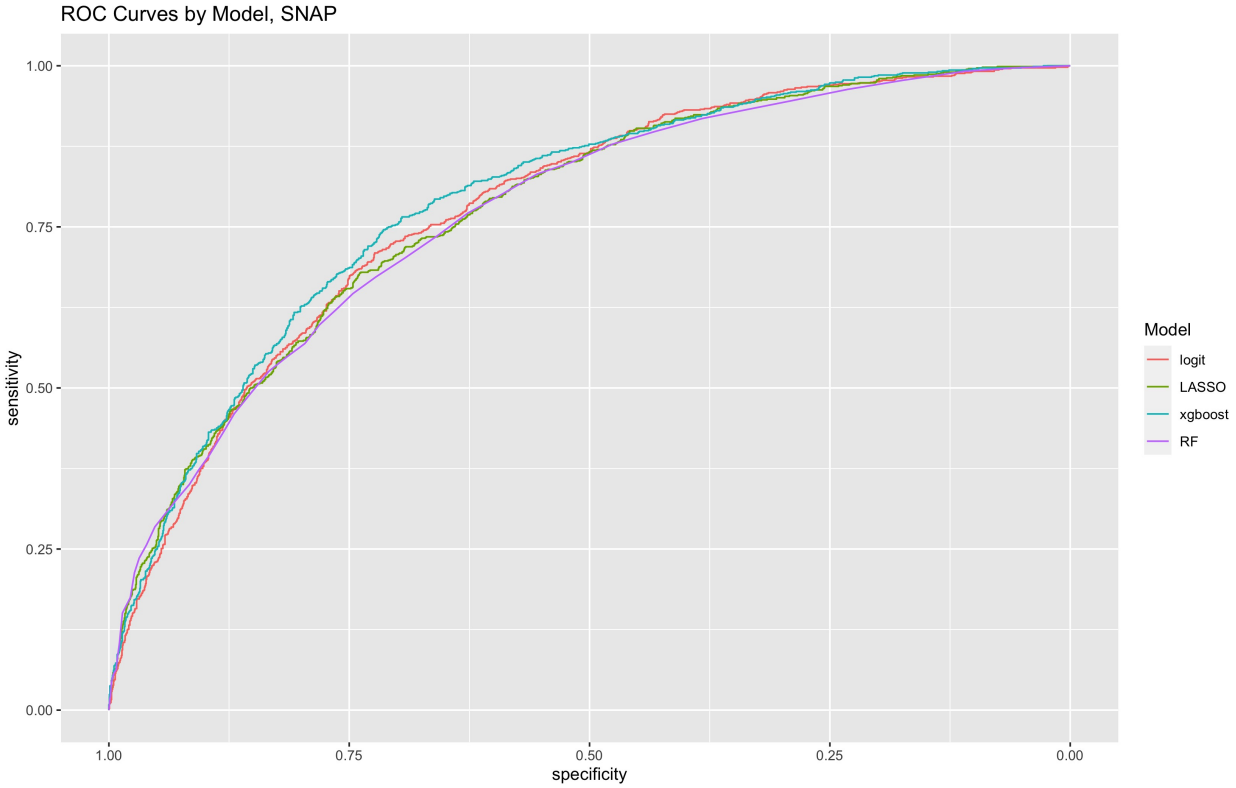
Note: Sensitivity (true positive rate) refers to the probability of a positive test, conditioned on truly being positive. Specificity (true negative rate) refers to the probability of a negative test, conditioned on truly being negative.

Another visualization of performance differences across models is the receiver operating characteristic (ROC) curve, which plots sensitivity and specificity of a model at various threshold settings. In effect, the ROC curve illustrates the attainable frontier of true positive and true negative rates for a specific model: the best model will (typically) produce better sensitivity at a given specificity than alternatives, and vice versa. Figure 1 presents ROC curves of the four models for predicting SNAP participation. While the difference is subtle,

[6]Training and tuning each model for predicting SNAP participation requires approximately 40, 190, and 55 seconds respectively for XGBoost, the elastic net, and random forests. However, note that XGBoost is tuned over a much larger grid of hyperparameters.

figure 1 shows that XGBoost has the best trade-off at the key range around the balanced accuracy of 0.72.

Figure 1: ROC Curves By Model, SNAP



Figures B1 through B3 present ROC curves for models predicting the other six welfare programs. For all seven programs, ROC curves for XGBoost are significantly above the diagonal, suggest that the models perform substantially better than a random classifier (the ROC curve of which would be a diagonal line). When comparing across models, XGBoost's performance is either the best, or not meaningfully worse than the best model, for six out of the seven programs. The only program where an alternate model has a clearly better area under the curve (AUC) measure is ACA, with LASSO being the best model.

The ROC curve comparison also highlights a common consideration for classification models: the trade-off between type 1 (false positive) and type 2 (false negative) errors. This is not a decision we make explicitly, as we impute data based on estimated probabilities of take-up, and assign participation up to the targeted take-up rate regardless of the actual

assigned likelihood of participation.[7] For policy purposes, however, it may be desirable to achieve a balanced performance across participants or non-participants, or focus on correctly identifying one type of eligible household relative to the other.

2.3 Adjusting for the Under-Reporting in Program Participation

After estimating the baseline take-up model, we need to correct CPS ASEC for the under-reporting of program take-up. Namely, for each program-eligible observation in ASEC that does not report participation, we assign participation indicators to match national statistics on take-up rates.

First, to assess the degree at which CPS ASEC underestimates households' participation in public assistance programs and tax credits we contrast the reported program take-up using observed data (see Table 1) to the "target" take-up rates that are based on the administrative data. Target take-up rates are reported in Table 3. Target rates are computed by dividing the enrollment data from various administrative data source on the number of eligible people estimated using the Policy Rules Database (Ilin and Terry 2021) applied to the 2019 CPS ASEC.

Results show that CPS ASEC indeed significantly underestimates take-up for all seven programs. The degree of underestimation varies by program. Child Tax Credit has the lowest difference between reported and actual take-up rates (6.5 percentage points). While, Medicaid for children has the largest discrepancy – the reported take-up rate is 38 percentage point lower that the one estimated using administrative data.

We address the discrepancy between reported and target take-up rates in the following way. First, we use the model from the previous step estimated using reported take-up to assign the probability of participation to each CPS ASEC unit that is eligible for a program but do not report the participation. Next, we impute participation to non-respondents with the highest predicted probability of participation until the overall take-up rate matches

[7] In other words, the only determinant for imputed participation is the relative likelihood among all eligible households.

Table 3: Reported and Target Take-Up Rate of Seven Public Assistance Programs

	Number of Participating Individuals ('000)	Number of Eligible Individuals ('000)	Target Take Up Rate (%)	Reported Take Up Rate (%)
SNAP	35,702	54,150	65.9	31.3
Housing Choice Voucher	5,249	46,559	11.3	7.8
Medicaid for Adults*	18,040	24,096	79.9	44.3
Medicaid for Children/CHIP**	29,232	34,917	83.7	55.7
ACA Subsidy	9,593	124,400	7.71	3.5
EITC	N/A	N/A	78.1	45.7
CTC	48,962	58,081	84.3	77.8

* Excluding dual Medicaid-Medicare enrollees and non-elderly adults with disabilities
** Excluding children with special needs care
Sources: Number of eligible individuals for each program are computed using the Policy Rules Database (Ilin and Terry 2021) applied to the 2019 Annual Social and Economic Supplement of the Current Population Survey. SNAP enrollment numbers are from SNAP Data Tables, Food and Nutrition Service, U.S. Department of Agriculture. Section 8 Housing Voucher enrollment data is from 2019 Picture of Subsidized Households, United States Department of Housing and Urban Development. Enrollment in Medicaid and CHIP is from Open Data, Center for Medicare and Medicaid Services; ACA Premium Subsidy enrollment is from 2019 Marketplace Open Enrollment Period Public Use Files, Center for Medicare and Medicaid Services. Estimates of the EITC take up is taken directly from the Internal Revenue Services. Number of tax returns with CTC is from Estimates of Federal Tax Expenditures for Fiscal Year 2019-2023, Joint Committee on Taxation

administrative numbers (target rates) reported in Table 3.

Finally, as a last step, we re-estimate the take-up probability model using the program take-up data adjusted to the target take-up rates to get the final estimates.

3 Application: Imputing Program Take-up for ACS respondents

Estimated model of program take-up has important practical applications. It allows researchers to predict out-of-sample probability of participation in social safety net programs to any datasets with detailed individual- or family- level demographic and income information. Examples of such datasets in the U.S. are Monthly Current Population Survey, Survey of Income and Program Participation, Panel Survey of Income Dynamics (PSID), and American Community survey (ACS).

In this section, we demonstrate how our take-up model can be applied to the 2019 American Community Survey (ACS) to study patterns of U.S. social safety net program participation in the United States. Namely, for each ACS household, we estimate program eligibility and impute participation for seven major public assistance programs and describe

five stylized facts about program take-up and participation in the U.S.

American Community Survey (ACS) is a representative survey of the U.S. population conducted annually that provides social, economic, housing and population data. ACS provides data for the nation, states, counties and other geographic areas down to the block group level. Out-of-sample prediction of program participation is conducted in two stages. First, we use estimates from our model to predict *probability* of participating in each public assistance program and tax credit for every non-elderly household in the 2019 ACS. At a second stage, we impute program participation to non-respondents with the highest predicted probability of participation until the overall take-up rate matches target take-up rates from Table 3.

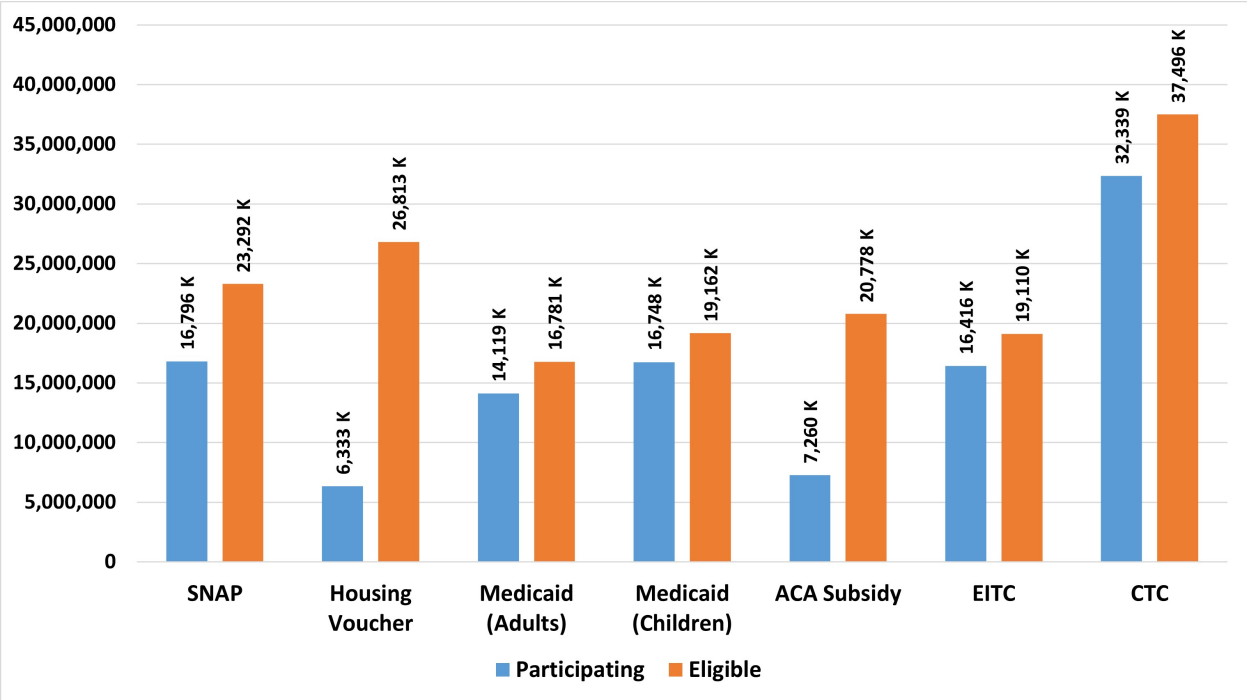
3.1 Program Eligibility and Participation

Firs, we estimate and report the number of eligible and participating households for each of seven major public assistance programs - SNAP, Housing Choice Voucher, Medicaid for Adults, Medicaid for Children/CHIP, ACA subsidy, EITC, and CTC. The results are reported in Figure 2. We estimate that the Child Tax Credit (CTC) has the highest number of eligible families (37,496 thousand) and also the largest number of participating families (32,339 thousand). Section 8 Housing Choice Voucher program grants eligibility to the second largest number of families (26,813 thousand), but has the lowest number of families that actually participate in the program (6,333 thousand). It is well known that Housing Voucher program has limited funding and as the results, eligible families must wait for years before they can get off the waitlist and start receiving the voucher (Acosta and Gartland 2021).

Medicaid for children has the largest participation among 19,162 thousand eligible families. Almost all families that are eligible for free public health insurance coverage for their children, participate in the program. This estimates are in line with the conclusion from the academic analysis that suggests that Medicaid and CHIP programs are very effective in decreasing the number of uninsured children in the nation (Kenney et al. 2012; Haley et al.

2018, 2019).

Figure 2: Estimated Number of Eligible and Participating Families by Program



Source: 2019 American Community Survey

3.2 Multiple Public Benefits Program Participation

Little is known about the extent to which low-income households receive multiple benefits and in what combinations. Some past studies have examined receipt of multiple benefits Acs and Loprest (2005); Rosenbaum (2011); Edelstein et al. (2014); Chien and Macartney (2019). Our methodology allows us to reassess this question and estimate the most common combinations of public assistance programs for non-elderly American families.

We start out by estimating the correlation between households' participation in seven public assistance programs of interest. Table 4 summarizes the results. We observe a strong correlation between participation in the ACA premium subsidy and participation in the Housing Choice Voucher program (0.96); between the participation in the Child Tax Credit and participation in the Housing Choice Voucher (0.69); between participation in the ACA

premium subsidy and participation in the Child Tax Credit (0.68); and between the participation in the Medicaid for Adults and Medicaid for Children (0.57). There is no statistically significant correlation between family's participation in Earned Income Tax Credit (EITC) and participation in Medicaid for Children. The correlation between participation in SNAP and Earned Income Tax Credit is estimated to be low (0.08) as well as the correlation between the participation in Child Tax Credit and in the Medicaid for Adults (0.06).

Table 4: Correlation Between Program Participation

	SNAP	Housing Voucher	Medicaid (Adults)	Medicaid (Children)	ACA Subsidy	EITC	CTC
SNAP	1						
Housing Voucher	0.29	1					
Medicaid (Adults)	0.42	0.24	1				
Medicaid (Children)	0.32	0.20	0.57	1			
ACA Subsidy	0.25	0.96	0.17	0.12	1		
EITC	0.08	0.29	0.06	-0.00	0.28	1	
CTC	0.12	0.69	0.08	0.06	0.68	0.36	1

Source: 2019 American Community Survey (ACS), authors' calculations

Next, we look at ten most common program combinations among U.S. non-elderly families. Results are reported in Table 5. We look separately at the most common benefits bundles for the population overall (Panel A) and for families with children (Panel B). Across both samples, Child Tax Credit is the most common benefit bundle – 18.5 percent of all non-elderly families and 18.3 percent of non-elderly families with children participate in this program.

For the population as a whole, SNAP, Housing Choice Voucher, public health insurance programs (Medicaid for Adults, Medicaid for Children, and ACA subsidy), and tax credits (EITC, CTC) is the second most common public benefits combination. However, although it is the second most common bundle, it is very uncommon – only 4.3 percent of families have this combination of programs. Medicaid for Children and CTC is the second most common bundle among families with children. But again, only 2.2 percent of families with children actually receive this bundle.

Table 5: Most Common Combinations of Public Assistance Benefits

Panel A: All Families

	Benefit Bundle	Number of Families	Share
1.	CTC	18,970,990	0.185
2.	SNAP, Section 8, Medicaid (Adults) Medicaid (Children) ACA EITC CTC	4,424,557	0.043
3.	Medicaid (Children), CTC	2,347,244	0.023
4.	SNAP, Medicaid (Adults)	2,066,497	0.020
5.	ACA	1,847,454	0.018
6.	Medicaid (Children), EITC, CTC	1,810,949	0.018
7.	SNAP	1,593,154	0.016
8.	SNAP, Medicaid (Adults), Medicaid (Children), EITC, CTC	1,363,334	0.013
9.	SNAP, Medicaid (Children), EITC, CTC	1,087,323	0.011
10.	SNAP, Medicaid (Adults), Medicaid (Children), EITC	1,067,402	0.010

Panel B: Families with Children

	Benefit Bundle	Number of Families	Share
1	CTC	18,682,927	0.183
2	Medicaid (Children), CTC	2,293,266	0.022
3	Medicaid (Children), EITC, CTC	1,755,775	0.017
4	SNAP, Medicaid (Adults), Medicaid (Children), EITC, CTC	1,272,327	0.012
5	SNAP, Medicaid (Children), EITC, CTC	1,032,068	0.010
6	SNAP, Medicaid (Adults), Medicaid (Children), EITC	965,264	0.009
7	Medicaid (Children)	713,515	0.007
8	Medicaid (Children), EITC	639,815	0.006
9	SNAP, Medicaid (Children), EITC	588,698	0.006
10	EITC, CTC	501,601	0.005

Source: 2019 American Community Survey (ACS), authors' calculations

3.3 Characteristics of Participants and Non-participants

Next, we look into the demographic differences between participating and non-participating eligible families in SNAP and EITC.

SNAP is the major public assistance program in the U.S., that is also known to have significant barriers to participation such as government-imposed administrative burden (Fox et al. 2019; Lopoo et al. 2020; Nicholson-Crotty et al. 2021) and stigma (Gaines-Turner et al. 2019; George-Lucas and Brandon 2022; Fairfax 2021). EITC is the most well know means tested tax credit for working families that is less stigmatized and is not generally associated administrative burden (Sykes et al. 2015). Because of these important differences in the level of stigma and administrative burden associated with SNAP and EITC we choose to analyse these two programs. Table 6 reports characteristics of participating and non-participating

families separately for each program.

Table 6: Characteristics of SNAP and EITC Participants and Non-Participants

	SNAP		EITC	
	Participants	Non-participants	Participants	Non-participants
Age	32.6	35.2	27.2	49.4
Annual Income ($)	24,654	60,769	34,291	23,947
Family Size	2.1	1.5	2.5	1.1
Number of Children	0.77	0.19	1.1	0.3
Residence: Metro	0.76	0.83	0.78	0.70
Married	0.08	0.10	0.12	0.10
Bachelor's degree or higher	0.07	0.31	0.09	0.14
White, non-Hispanic	0.49	0.59	0.46	0.62
Black, non-Hispanic	0.23	0.10	0.20	0.20
Hispanic	0.20	0.18	0.24	0.09

Source: 2019 American Community Survey (ACS), authors' calculations

Eligible SNAP families that choose to participate tend to have lower income, larger family size and greater number of kids. Significantly higher share of SNAP participants do not have a college degree. Slightly higher share of SNAP participants are single and live in rural areas. Finally, a larger share of participating families are Black and Hispanic relative to non-participants.

Participants in EITC are much younger than non-participants, have higher income (that potentially gives them higher dollar value of the credit) and have more children (significantly larger credit). EITC participants are less likely to be White and significantly more likely to be Hispanic. Slightly larger share of EITC participants are married, and without college degree.

3.4 Variation in Program Take-Up by State

Finally, using the example of SNAP , we demonstrate how the program take-up varies across the U.S. states. Figure 3 plots SNAP eligibility rates (share of all non-elderly families in a state that participate are eligible for the program), SNAP participation rate (share of all non-elderly families in a state that participate in the program), and SNAP take-up rates

(share of eligible families that chose to participate) by state.

SNAP eligibility rate, participation rate, and take-up rates vary substantially by state. Participation rates vary from 11 percent of the population in California to 25 percent of state population in West Virginia. Difference in participation rates can be partially explained by the variation in the share of eligible families. West Virginia has much lower income than California, and thus higher number of eligible families.

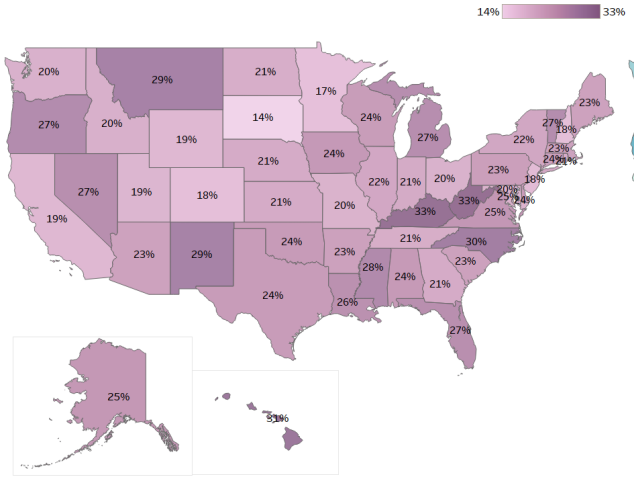
Eligibility rates vary from 14 percent in South Dakota to 33 percent in Kentucky. This variation is a combination of differences in average incomes across states and a differences in the eligibility limits. Take-up rates vary from 60 percent in California to 87 percent in South Dakota. Variation in SNAP take-up rates can be explained in the differences in demographic characteristics of families eligible for the program (see Table 6). Note, that our estimates of SNAP take up differ significantly from the Food and Nutrition Services (FNS) estimates. FNS estimates define eligible households as those that pass all applicable Federal SNAP income and resource tests. This definition does not account for the expanded expanded income eligibility rules under the Broad Based Categorical Eligibility. Thus, FNS estimates significantly underestimate the share of families eligible for SNAP and overestimate the take-up rates.

4 Conclusion & Discussion

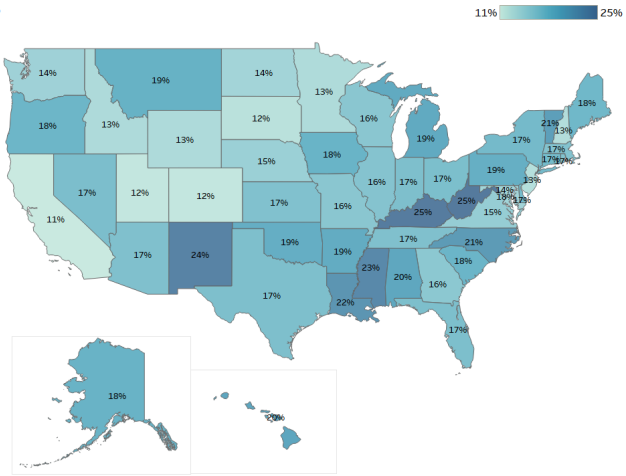
In this paper we develop and implement a novel machine-learning method to estimate the probability of programs take-up among eligible people for seven major U.S. social safety net programs: Supplemental Nutrition Assistance Program (SNAP), Section 8 Housing Choice Voucher Program, Medicaid for Adults, Medicaid for Children/Children Health Insurance Program (CHIP), Affordable Care Act (ACA) marketplace health insurance subsidies, Earned Income Tax Credit (EITC), and Child Tax Credit (CTC). Our machine-learning approach does not rely on linked administrative data, but rather uses survey data itself to

Figure 3: SNAP Eligibility, Participation, and Take-Up by State

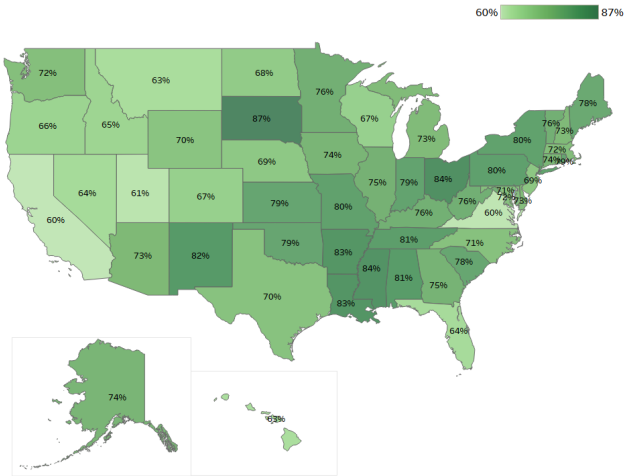
(a) SNAP Eligibility Rates by State



(b) SNAP Participation Rates by State



(c) SNAP Take-Up Rates by State



Source: 2019 American Community Survey, authors' calculations

determine the predominating demographic factors associated with the take up of public assistance benefits among eligible families, and subsequently estimate the probability of program participation given those factors.

We evaluate and contrast four alternative binary classification algorithms to estimate the probability of participation (gradient-boosted trees, random forests, and logistic regression with and without penalization). We favor XGboost (Chen and Guestrin 2016), a gradient-boosted classifier, for its relative efficiency and high performance of 70 to 90 percent out-of-sample accuracy. To account for the pervasive underreporting of program receipt, each

program-eligible, but non-reporting CPS ASEC observation is assigned a participation indicator that aligns with national statistics on take-up for a given program. After adjusting for underreporting of program receipt, our model is then re-estimated with the adjusted ASEC data to get final estimates.

Our method to impute program participation has important practical applications for economics and public policy researchers. Model estimates can be easily applied to any individual- or family-level dataset with detailed demographic and income information. We demonstrate how the model works by applying it to the 2019 American Community Survey (ACS). We establish five stylized facts about participation in the U.S. social safety net programs.

First, among all U.S. social safety net programs, the Child Tax Credit (CTC) has by far the largest number of eligible and participating families. Housing Choice Voucher Program has the second largest number of eligible families, but the lowest uptake. Second, there is a strong correlation between participation in the ACA marketplace health insurance subsidies and in the Housing Choice Voucher program. There is no statistically meaningful correlation between participation in Earned Income Tax Credit (EITC) and Medicaid for Children/CHIP. Third, Child Tax Credit as a single benefit is the most common benefit bundle. Combination of SNAP, Housing Choice Voucher, Medicaid for Adults, Medicaid for Children, ACA marketplace health insurance subsidies, EITC, and CTC is the second most common bundle, although only 4.3 percent of all families are estimated to receive it. Fourth, there is no clear relationship between demographic characteristics and program take up that holds across all programs. We find that this relationship is very program specific. Fifth, program eligibility, participation, and take-up rates vary significantly by state. For example, take-up rates of Supplemental Nutrition Assistance Program (SNAP) vary from 60 percent in California to 87 percent in South Dakota; SNAP eligibility rates vary from 14 percent in South Dakota to 33 percent in Kentucky.

Our novel machine learning-based technique developed in this paper, combined with the

established stylized facts about social programs take-up, can inform future research which is thought to better understand the determinants of social safety net program participation.

References

Acosta, S. and Gartland, E. (2021). Families wait years for housing vouchers due to inadequate funding. page 12.

Acs, G. and Loprest, P. (2005). Who are low-income working families? washington, dc: Urban institute.

Anderson, P. M. and Meyer, B. D. (1997). Unemployment insurance takeup rates and the after-tax value of benefits. *The quarterly journal of economics*, 112(3):913–937.

Bee, A. and Mitchell, J. (2017). Do older americans have more income than we think? In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, volume 110, pages 1–85. JSTOR.

Center on Budget and Policy Priorities (2022). Where do our federal tax dollars go?

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chien, N. (2015). Estimates of child care eligibility and receipt for fiscal year 2012. *Parameters*, 14(2,177,000):15.

Chien, N. and Macartney, S. (2019). What happens when people increase their earnings? effective marginal tax rates for low-income households. *Office of the Assistant Secretary for Planning & Evaluation, US Department of Health & Human Services. Brief*, 2.

Coder, J. and Scoon-Rogers, L. (1996). Evaluating the quality of income data collected in the annual supplement to the march current population survey and the survey of income and program participation. Technical report, Citeseer.

Edelstein, S., Pergamit, M. R., and Ratcliffe, C. (2014). Characteristics of families receiving multiple public benefits. *Washington, DC: Urban Institute. http://www. urban. org/research/publication/characteristics-families-receiving-multiple-public-benefits*.

Fairfax, P. (2021). Snap and stigma: Analyzing the food stamps program.

Foster, A. C. and Rojas, A. (2018). Program participation and spending patterns of families receiving government means-tested assistance. *Monthly Labor Review*, pages 1–22.

Fox, A., Feng, W., and Reynolds, M. (2019). The effect of administrative burden on state safety-net participation: Evidence from snap, tanf and medicaid. *Public Administration Review*.

Gaines-Turner, T., Simmons, J. C., and Chilton, M. (2019). Recommendations from snap participants to improve wages and end stigma. *American journal of public health*, 109(12):1664–1667.

George-Lucas, D. and Brandon, P. D. (2022). Explaining snap nonparticipation among food-insecure poor and near-poor households with dependent children. *Journal of Poverty*, 26(6):483–499.

Giannarelli, L. (2019). What was the tanf participation rate in 2016?

Haley, J., Kenney, G. M., Wang, R., Pan, C., Lynch, V., and Buettgens, M. (2019). Improvements in uninsurance and medicaid/chip participation among children and parents stalled in 2017. *Washington, DC: Urban Institute*.

Haley, J. M., Kenney, G. M., Wang, R., Lynch, V., and Buettgens, M. (2018). Medicaid/chip participation reached 93.7 percent among eligible children in 2016. *Health Affairs*, 37(8):1194–1199.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, pages 278–282, M. IEEE Computer Society.

Ilin, E. and Terry, E. (2021). The Policy Rules Database.

Jolliffe, D., Gundersen, C., Tiehen, L., and Winicki, J. (2005). Food stamp benefits and child poverty. *American Journal of Agricultural Economics*, 87(3):569–581.

Kenney, G., Lynch, V., Huntress, M., Haley, J., and Anderson, N. (2012). Medicaid/chip participation among children and parents. *Timely Analysis of Health Policy Issues*.

Kosar, G. and Moffitt, R. A. (2017). Trends in cumulative marginal tax rates facing low-income families, 1997–2007. *Tax Policy and the Economy*, 31(1):43–70.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283.

Lopoo, L. M., Heflin, C., Boskovski, J., et al. (2020). Testing behavioral interventions designed to improve on-time snap recertification. *Journal of Behavioral Public Administration*, 3(2).

Lynch, V., Resnick, D. M., Staveley, J., and Taeuber, C. (2008). *Differences in estimates of public assistance recipiency between surveys and administrative records*. Jacob France Institute, University of Baltimore.

Meyer, B. D. and Mittag, N. (2019). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics*, 11(2):176–204.

Meyer, B. D., Mittag, N., and Goerge, R. M. (2022). Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. *Journal of Human Resources*, 57(5):1605–1644.

Meyer, B. D., Mok, W. K., and Sullivan, J. X. (2009). The under-reporting of transfers in household surveys: its nature and consequences. Technical report, National Bureau of Economic Research.

Meyer, B. D., Mok, W. K., and Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226.

Meyer, B. D. and Wu, D. (2018). The poverty reduction of social security and means-tested transfers. *ILR Review*, 71(5):1106–1153.

Nicholson-Crotty, J., Miller, S. M., Keiser, L. R., et al. (2021). Administrative burden, social construction, and public support for government programs. *Journal of Behavioral Public Administration*, 4(1).

Riphahn, R. T. (2001). Rational poverty or poor rationality? the take-up of social assistance benefits. *Review of income and wealth*, 47(3):379–398.

Roemer, M. (2000). Assessing the quality of the march current population survey and the survey of income and program participation income estimates. *Unpublished paper, US Census Bureau*.

Rosenbaum, S. (2011). *Improving the delivery of key work supports: Policy & practice opportunities at a critical moment*. Center on Budget and Policy Priorities.

Rothbaum, J. L. (2015). Comparing income aggregates: How do the cps and acs match the national income and product accounts, 2007-2012. *US Census Bureau, SEHSD Working Paper*, 1.

Sykes, J., Križ, K., Edin, K., and Halpern-Meekin, S. (2015). Dignity and dreams: What the earned income tax credit (eitc) means to low-income families. *American Sociological Review*, 80(2):243–267.

Tibshirani, R. (1996). Regression shinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288.

Wheaton, L. (2016). Underreporting of means-tested transfer programs in the cps and sipp.

Yaniv, G. (1997). Welfare fraud and welfare stigma. *Journal of Economic Psychology*, 18(4):435–451.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Appendices

Appendix A Policy Rules Database

To estimate the eligibility and dollar value of public assistance programs for each individual in our sample, we use state-specific tax rules and transfer-payment eligibility rules as collected in the Policy Rules Database (Ilin and Terry 2021). The Policy Rules Database contains all of the major federal and state assistance programs, tax rules, and tax credits available to working adults and their dependents. Table A1 provides the full list of programs included in the Policy Rules Database.

Table A1: List of Taxes and Public Assistance Programs Included in the Policy Rules Database

Taxes and Tax Credits	
Personal Income Tax	Federal, State
Sales Tax	State
Federal Insurance Contribution Act (FICA) Tax	Federal
Earned Income Tax Credit (EITC)	Federal, State
Child Tax Credit (CTC)	Federal, State
Child and Dependent Care Tax Credit (CDCTC)	Federal, State
Public Assistance Programs	
School Breakfast Program (SBP) and National School Lunch Programs (NSLP)	Federal
Supplemental Nutrition Assistance Program (SNAP)	Federal, State
Special Supplemental Nutrition Program for Women, Infants, and Children (WIC)	Federal
(In Progress) Temporary Assistance for Needy Families (TANF)	State
Medicaid	Federal, State
Children's Health Insurance Program (CHIP)	Federal, State
Health Insurance Marketplace Subsidies	Federal
Employer Sponsored Health Insurance	State
The Housing Choice Voucher Program (Section 8)	State, County
(In Progress) Low Income Home Energy Assistance Program (LIHEAP)	State
State Voluntary Pre-K	State
Head Start/Early Head Start	Federal
Child Care and Development Fund (CCDF) Subsidies	State, County
Supplemental Security Income (SSI)	Federal, State
Social Security Disability Insurance (SSDI)	Federal

Source: The Policy Rules Database (Ilin and Terry 2021).

The Policy Rules Database is available at
https://github.com/FRB-Atlanta-Advancing-Careers/policy-rules-database

Appendix B Model Tuning

Table B1: Comparison of OOS Performance of Alternate Imputation Models, Section 8

	Logit	Elastic Net	XGBoost	Random Forest
Accuracy	0.915	0.916	0.924	0.924
Sensitivity	0.999	0.995	0.999	0.999
Specificity	0.011	0.038	0.047	0.012

Table B2: Comparison of OOS Performance of Alternate Imputation Models, Medicaid Adult

	Logit	Elastic Net	XGBoost	Random Forest
Accuracy	0.741	0.744	0.736	0.736
Sensitivity	0.825	0.702	0.699	0.690
Specificity	0.653	0.787	0.776	0.783

Table B3: Comparison of OOS Performance of Alternate Imputation Models, Medicaid Child

	Logit	Elastic Net	XGBoost	Random Forest
Accuracy	0.738	0.739	0.755	0.748
Sensitivity	0.745	0.643	0.680	0.643
Specificity	0.733	0.801	0.804	0.813

Table B4: Comparison of OOS Performance of Alternate Imputation Models, ACA

	Logit	Elastic Net	XGBoost	Random Forest
Accuracy	0.959	0.957	0.959	0.955
Sensitivity	0.999	0.998	0.997	0.997
Specificity	0.030	0.059	0.125	0.068

Table B5: Comparison of OOS Performance of Alternate Imputation Models, EITC

	Logit	Elastic Net	XGBoost	Random Forest
Accuracy	0.955	0.966	0.952	0.957
Sensitivity	0.948	0.948	0.924	0.921
Specificity	0.960	0.977	0.968	0.981

Table B6: Comparison of OOS Performance of Alternate Imputation Models, CTC

	Logit	Elastic Net	XGBoost	Random Forest
Accuracy	0.976	0.980	0.983	0.983
Sensitivity	0.052	0.011	0.000	0.000
Specificity	0.996	1.000	1.000	1.000

Figure B1: ROC Curves By Model, Section 8 and Medicaid Adult

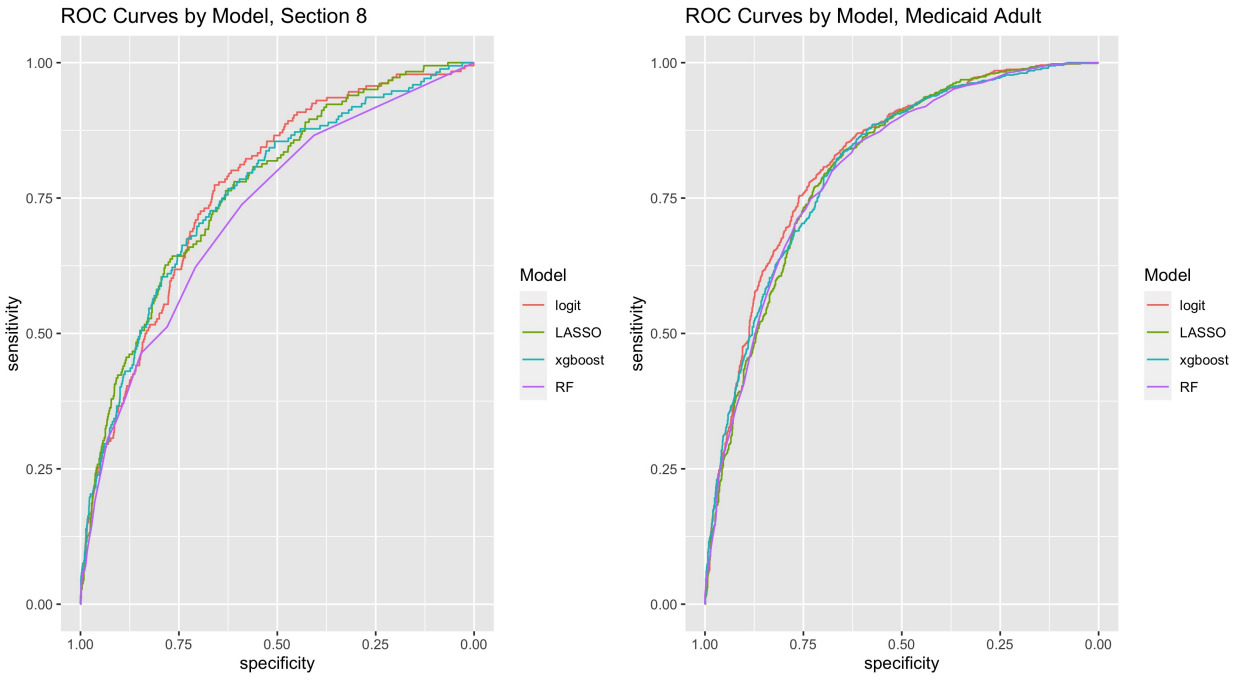


Figure B2: ROC Curves By Model, Medicaid Child and ACA

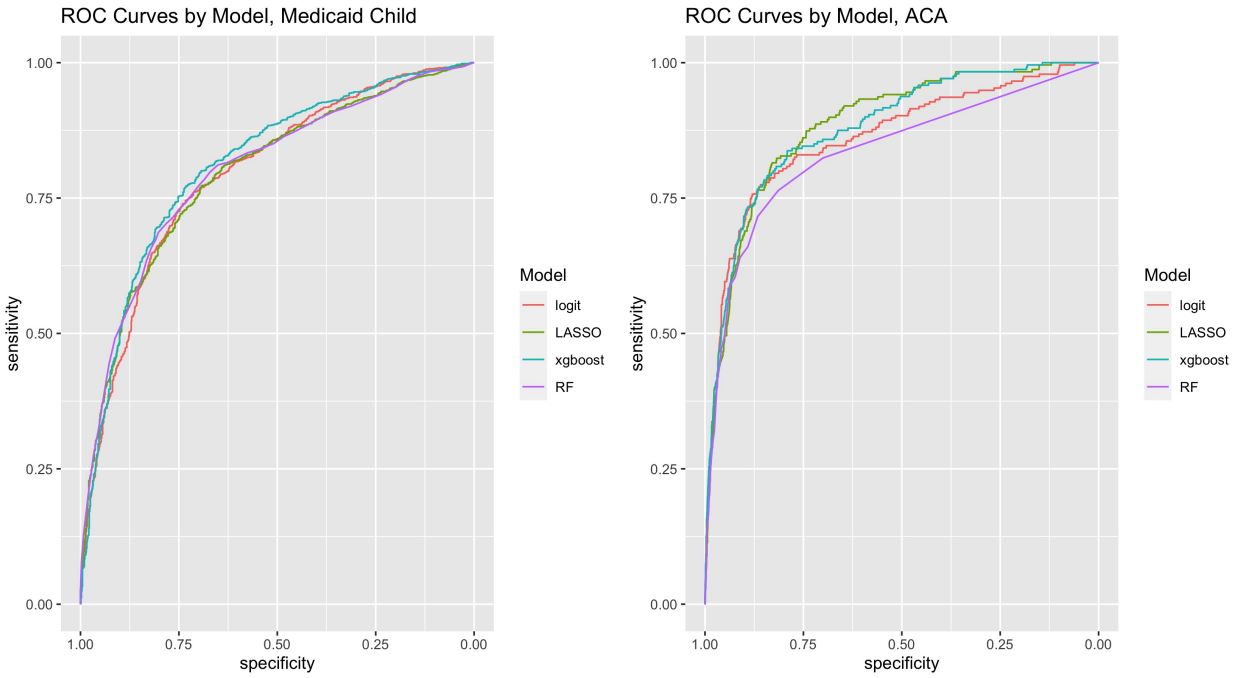


Figure B3: ROC Curves By Model, EITC and CTC