

Exploring the Application of Machine Learning Techniques to Construct R-indicators

Arcenis Rojas &
Lucilla Tan

Division of Consumer Expenditure Surveys
2019 AAPOR Conference
May 17, 2019

Consumer Expenditure Survey (CE)



Motivation

- **Problem:** Non-response bias violates many assumptions that are made during the sampling procedure and can lead to biased survey estimates.
- **Potential Solution:** Develop an indicator of representativeness of the respondent pool while data collection is still on-going.
 - ▶ This can inform the allocation of recruitment resources of under-represented groups.

Motivation

- Develop a representativeness indicator for a specific expenditure category (food) to monitor the representativeness of the respondent pool with respect to select characteristics over the course of data collection during the survey year.
 - ▶ **We're looking for variables that are associated with both food expenditures and survey participation.**

Background

Representativeness Indicator (*R-Indicator*):
Measures the risk of potential non-response bias based on weighted, estimated propensities of response.

$$R(\rho_x) = 1 - 2 * S(\hat{\rho}_i)$$

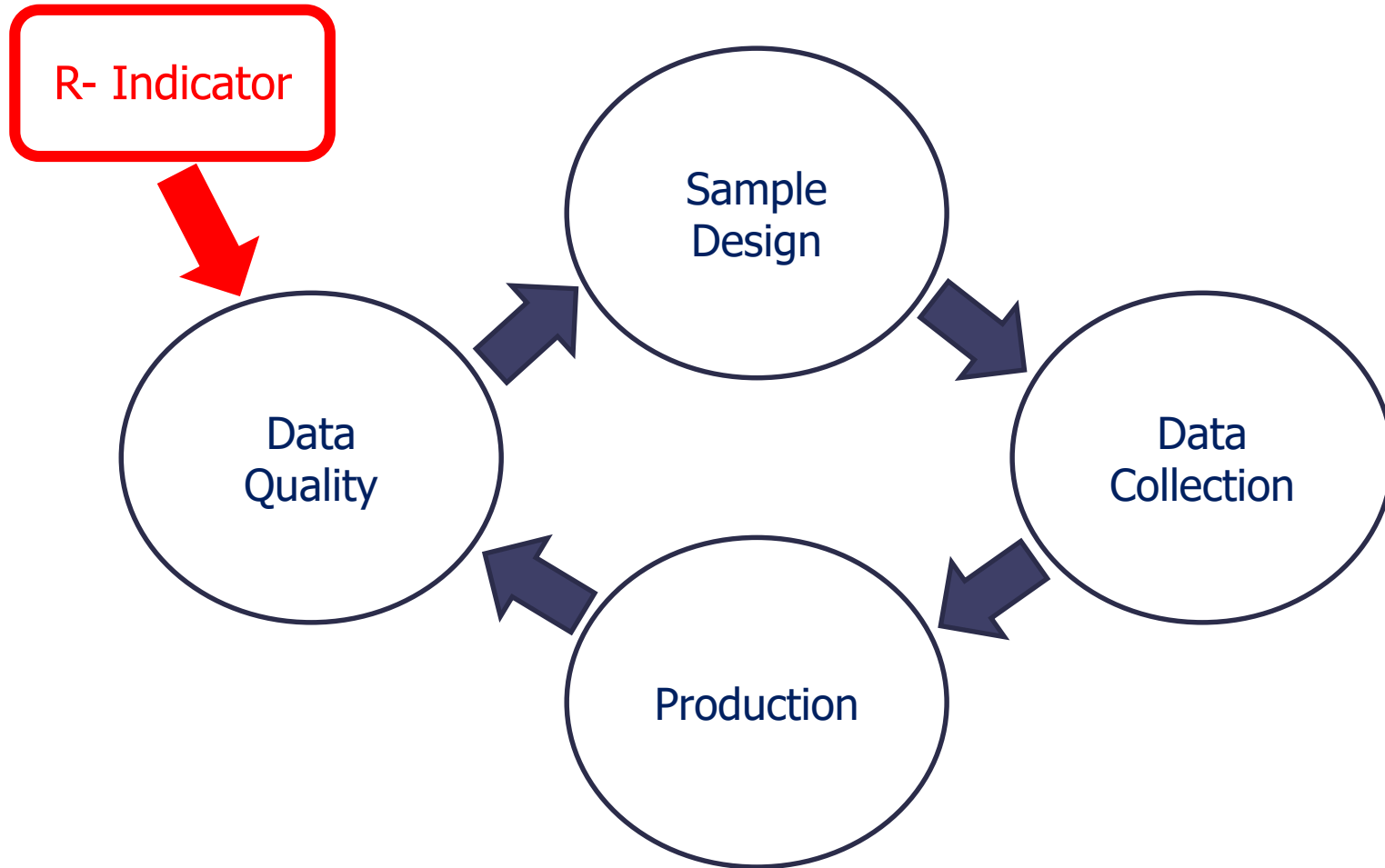
$$s(\hat{\rho}_i) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N d_i (\hat{\rho}_i - \hat{\rho})^2}$$

d_i : design weight for sample unit i

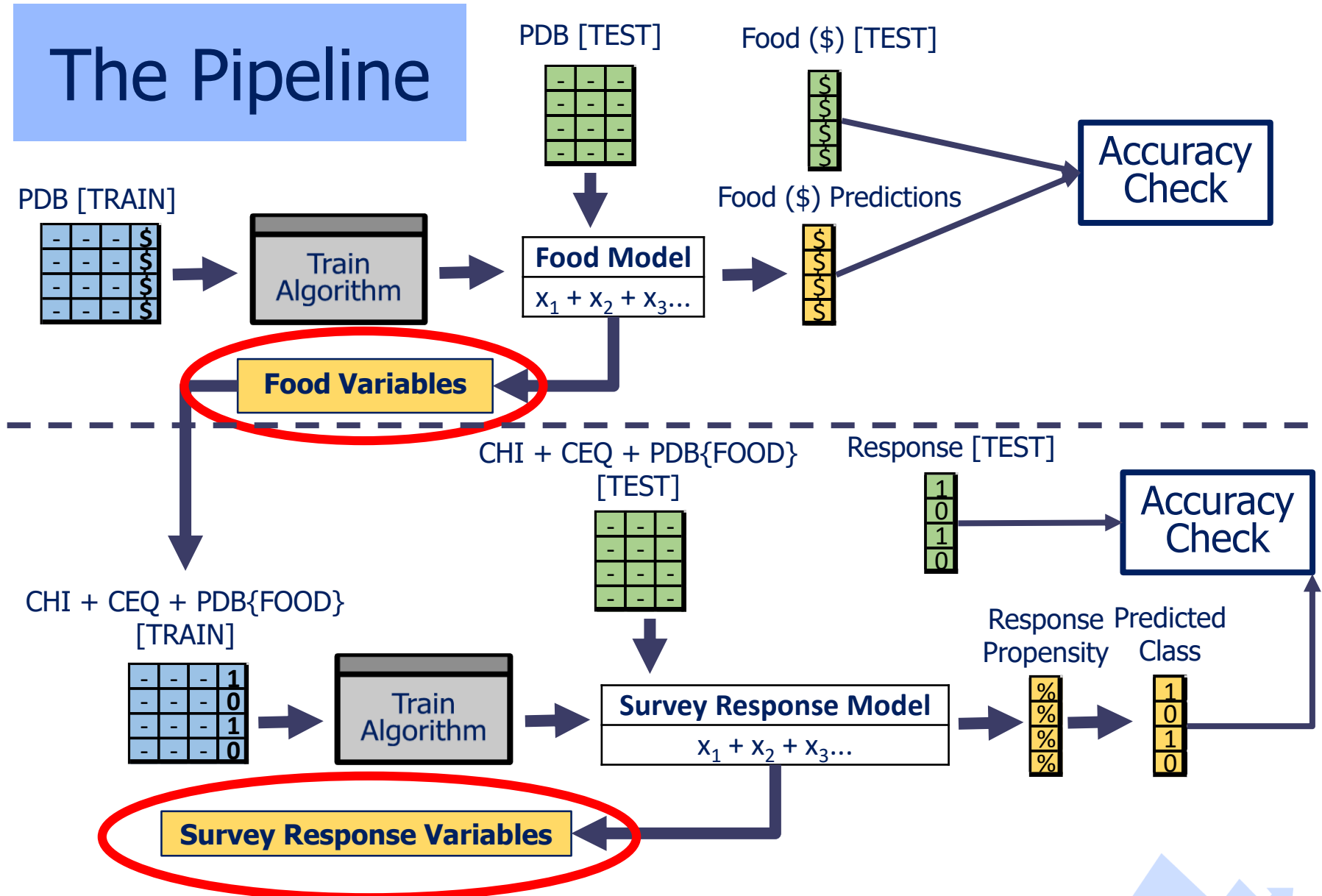
$\hat{\rho}$: weighted sample average of response propensities

$\hat{\rho}_i$: the estimated response propensity for unit i

Background



The Pipeline



Data Description

- **2015 Consumer Expenditure Interview Survey (CEQ)**
- **CE Interview Survey Contact History Instrument (CHI)**
- **The 2015 Census Planning Database (PDB)**
 - ▶ Geographic aggregation: tract-level (2010 Census boundaries)
 - ▶ Incorporates the 2009-2013 American Community Survey (ACS) five-year estimates
 - ▶ Latest data available at the time study started

Data Description

- Initial sample size: 36,226
 - ▶ After data cleaning: 32,255
 - ▶ 21,546 (66.8%) were survey participants
- Initial number of PDB variables: 114
 - ▶ After eliminating highly correlated and near-zero variance PDB variables: 54
- CHI variables: 2
- CEQ variables: 6

Data Description

- Data covered 5 periods over 2015
 - ▶ Feb
 - ▶ Feb-Mar
 - ▶ Feb-Jun
 - ▶ Feb-Sept
 - ▶ Feb-Dec
- We could compute an R-Indicator over time
 - ▶ Continuous monitoring is the motivation to build a repeatable process



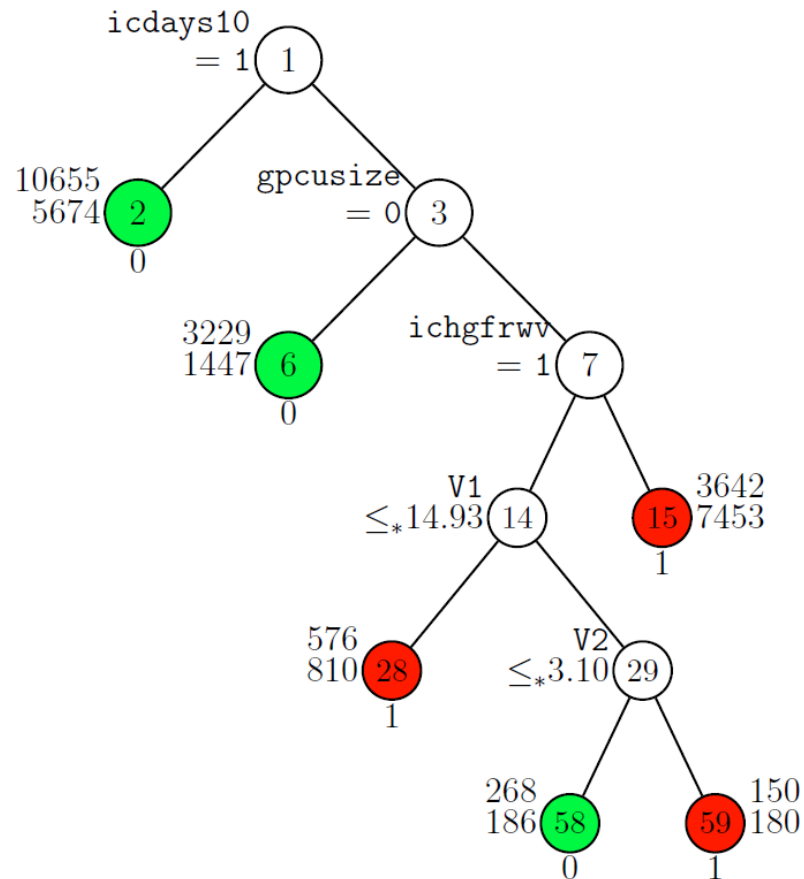
Selecting Algorithms

- Desired model characteristics:
 - ▶ High prediction accuracy
 - ▶ Dimension reduction
 - ▶ Interpretability
 - ▶ Smooth propensity distribution

Selecting Algorithms

	Classification Tree	Random Forest	Logistic Regression	LASSO
Prediction Accuracy				
Dimension Reduction				
Interpretability				
Smooth Propensity Distribution				

Classification Tree



Classification Tree


Pros:

- Easy to interpret
- Good dimension reduction

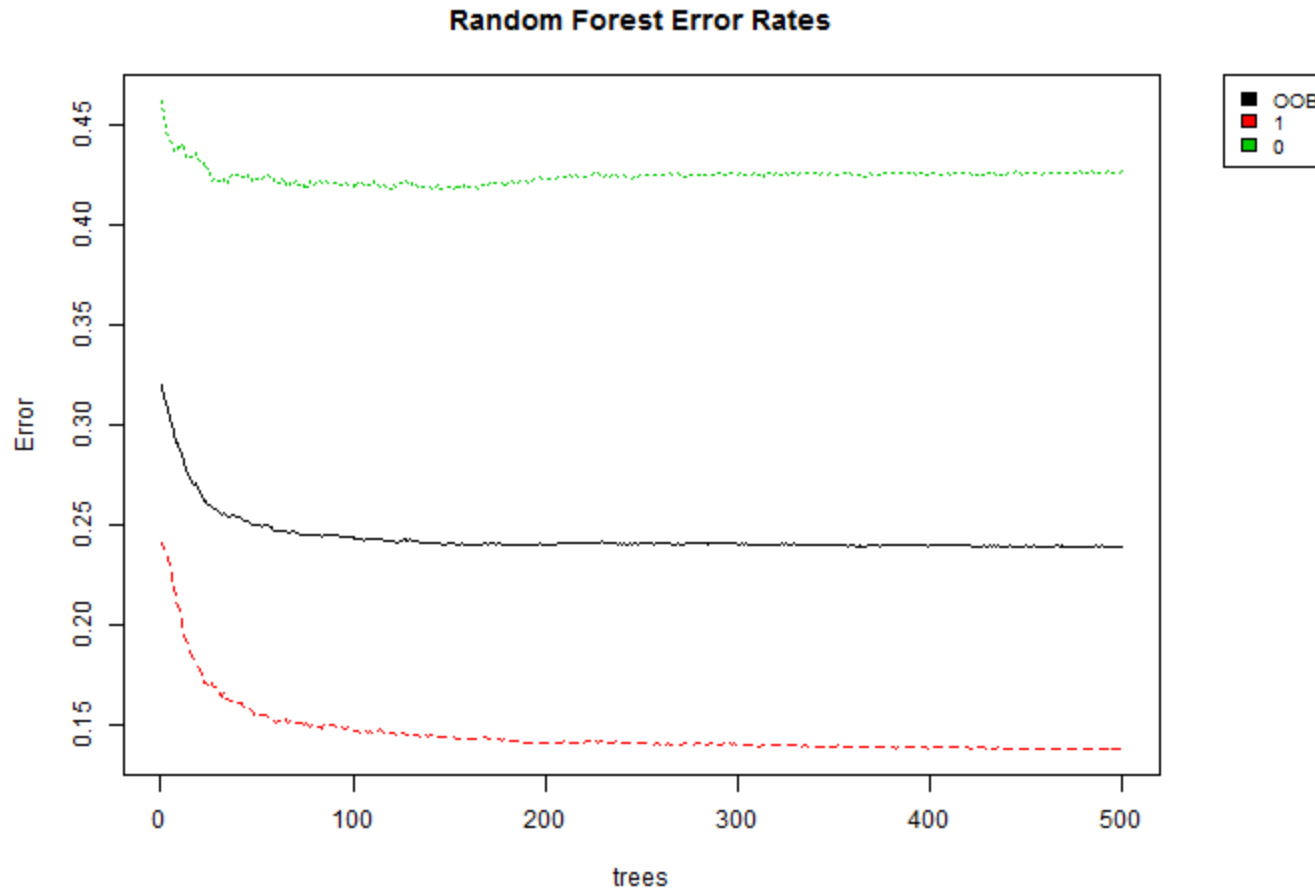
Cons:

- Propensities are “chunky”
- Cannot always handle missing values
- Sensitive to tuning parameter specification

Selecting Algorithms

	Classification Tree	Random Forest	Logistic Regression	LASSO
Prediction Accuracy				
Dimension Reduction				
Interpretability				
Smooth Propensity Distribution				

Random Forest



Random Forest

Pros:

- High accuracy
- Great for dimension reduction

Cons:

- Interpretation not as clear as other models
- Easily biased if not properly tuned
- Cannot handle missing values

Selecting Algorithms

	Classification Tree	Random Forest	Logistic Regression	LASSO
Prediction Accuracy	✓	✓		
Dimension Reduction	✓	✓		
Interpretability	✓	✗		
Smooth Propensity Distribution	✗	✓		

Logistic Regression

Pros:

- Easy to interpret
- Propensities are smooth
- Good for explaining variance

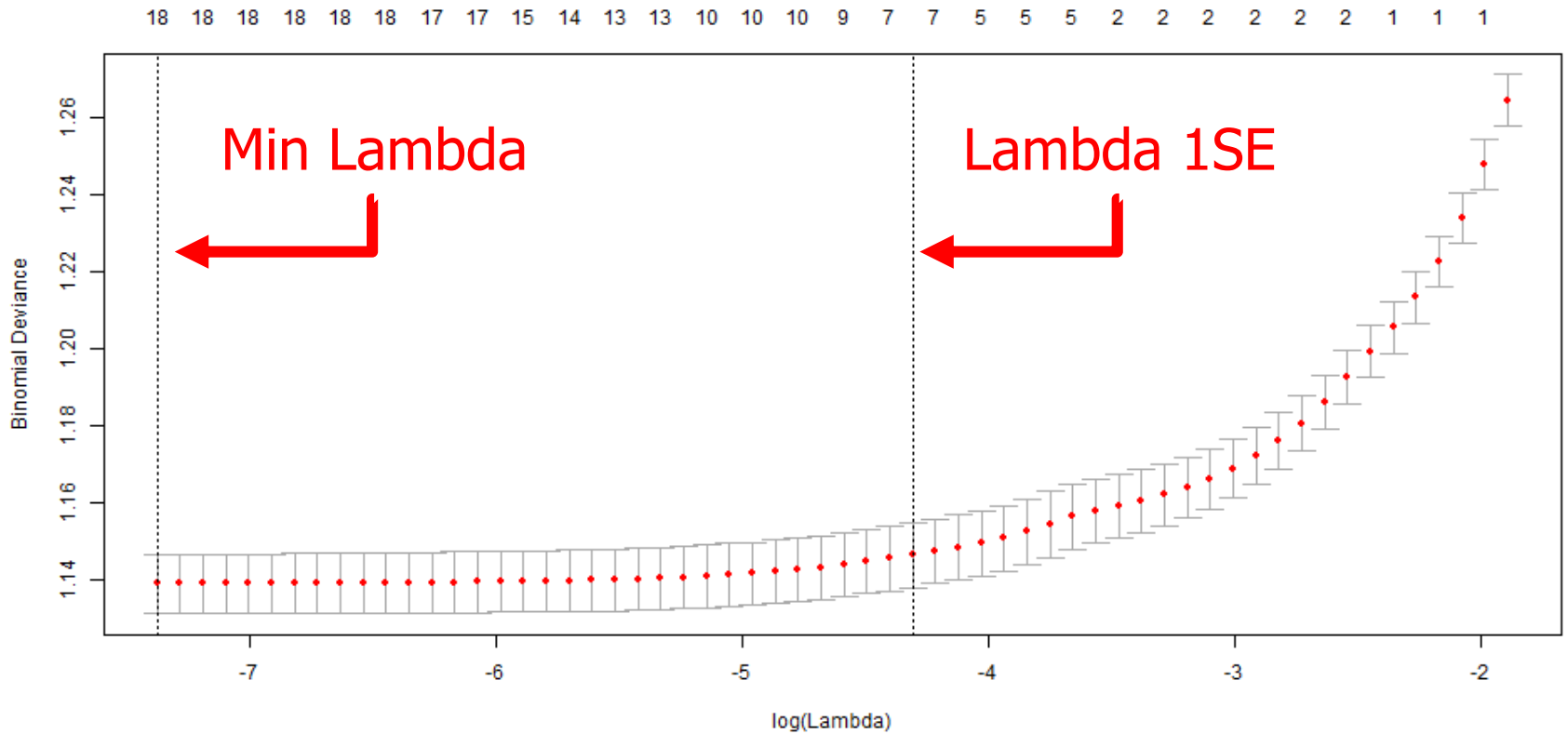
Cons:

- No dimension reduction
- Does not necessarily predict well

Selecting Algorithms

	Classification Tree	Random Forest	Logistic Regression	LASSO
Prediction Accuracy	✓	✓	✓	
Dimension Reduction	✓	✓	✗	
Interpretability	✓	✗	✓	
Smooth Propensity Distribution	✗	✓	✓	

LASSO



Least Absolute Shrinkage and Selection Operator (LASSO)

Pros:

- High prediction accuracy
- Easy interpretation
- Great for dimension reduction

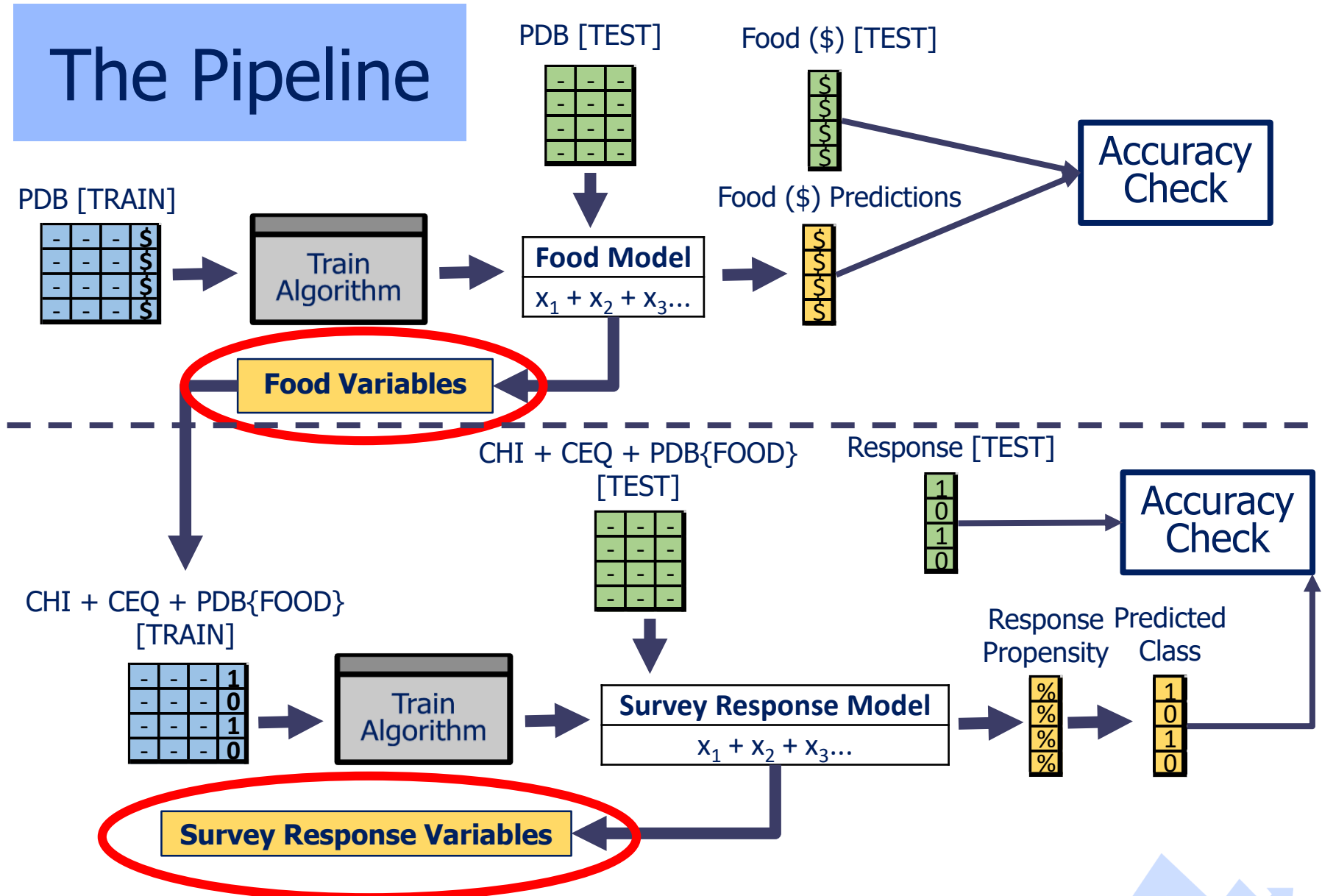
Cons:

- Coefficients do not necessarily indicate the magnitude of an effect.
- Narrative may not be intuitive

Selecting Algorithms

	Classification Tree	Random Forest	Logistic Regression	LASSO
Prediction Accuracy	✓	✓	✓	✓
Dimension Reduction	✓	✓	✗	✓
Interpretability	✓	✗	✓	✓
Smooth Propensity Distribution	✗	✓	✓	✓

The Pipeline



LASSO Explained

The lasso regression coefficient estimates are obtained by solving the optimization problem that can be generally characterized as:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} + \lambda \left(\sum_{j=1}^p |\beta_j| \right)$$

where $\lambda \geq 0$ is the shrinkage parameter that controls the relative impact of the two terms. The effect of the penalty is to get a more generalized (than a strict) fit to the data to minimize overfitting.

Training the LASSO

- Split the data 50-50 into a training set and a testing set
- Used 10-fold cross validation to find the best shrinkage parameter (λ) and used a mixing parameter of $\alpha = 1$
 - ▶ Selected the largest lambda within one standard error of the minimum cross-validation mean-standard error, which we call “Lambda 1SE”
- Ran LASSO

Predictors of Food Expenditure

- Covariates for Food Expenditure:
 - ▶ Started with 54 variables from the PDB
 - ▶ After running LASSO with “Lambda 1SE” we were left with 2 variables with non-zero coefficients:
 - Average Household Income (PDB)
 - Average House Value (PDB)

Predicting Survey Participation

■ Inputs:

- Average Household Income (PDB)
- Average House value (PDB)
- Census Region (CEQ)
- Dwelling Unit Structure Type (CEQ)
- Household Size (CEQ)
- Homeowner / Renter (CEQ)
- Urbanicity (CEQ)
- Survey Wave (CEQ)
- Number of Contact Attempts (CHI)
- Ever Changed Interviewer (CHI)

Predicting Survey Participation

- After running LASSO with “Lambda 1SE” we were left with 6 variables with non-zero coefficients:
 - Average Household Income (PDB)
 - Household Size (CEQ)
 - Homeowner / Renter (CEQ)
 - Urbanicity (CEQ)
 - Number of Contact Attempts (CHI)
 - Ever Changed Interviewer (CHI)

Model Accuracy

- Baseline was the unregularized GLM Logistic model

Unit response model: model performance comparison using Test subsample with model parameters estimated on Train subsample

Model predictors	Regression	Proportion prediction accuracy (cut-off value prob >0.5)*	Area under the ROC
λ 1SE- regularized (6 predictors)	GLM logistic	0.7261	0.699
Unregularized (10 predictors)	GLM logistic	0.7263	0.703

* Units with predicted probabilities >0.5 were classified as respondents.

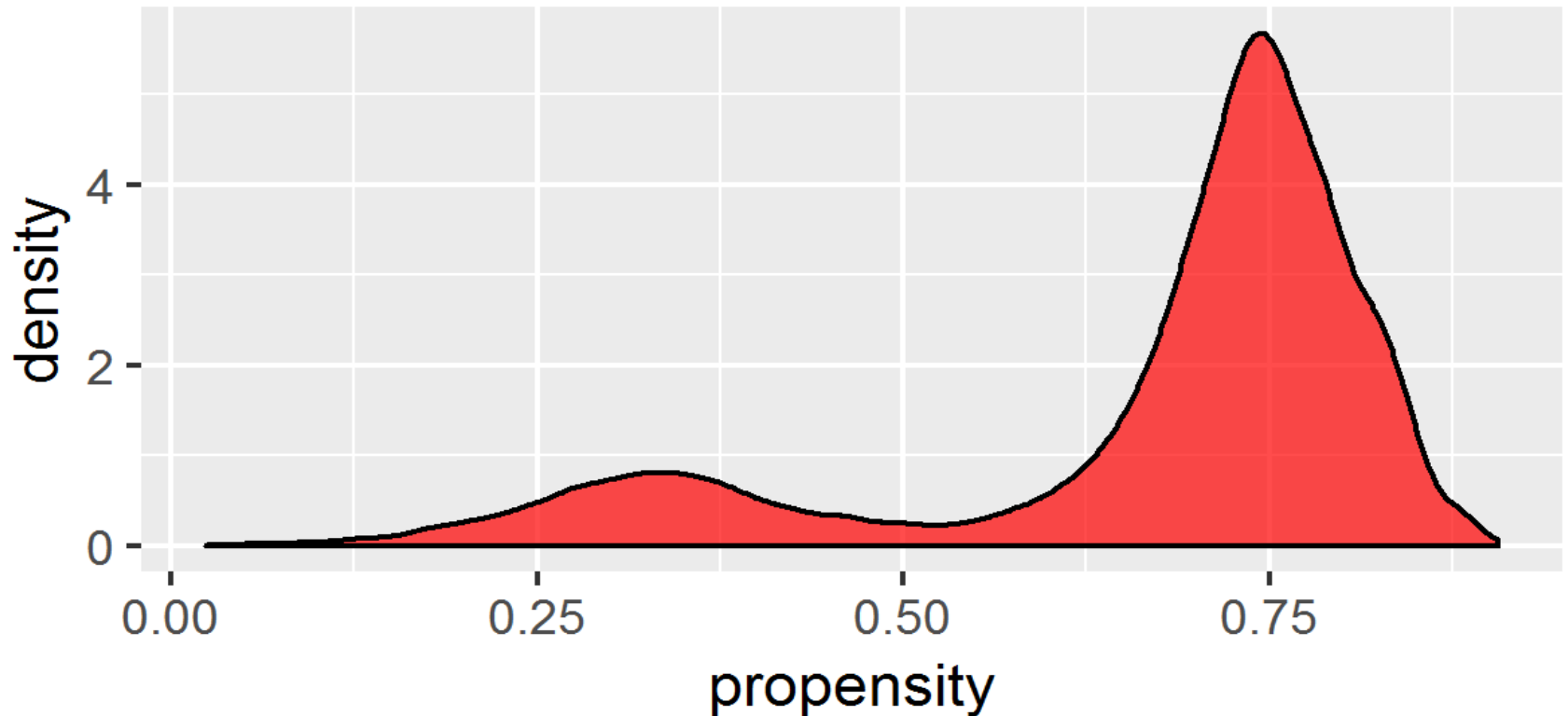
Period 5 Final Model

* Prediction Accuracy = 72.8%

Predictors	Coeff	SE	p-value
(Intercept)	1.414	0.041	0.000
Household income	-2.79E-06	3.49E-07	0.000
No. contact attempts	-0.089	0.004	0.000
HH size - one	-0.088	0.031	0.005
HH size - three	0.369	0.037	0.000
HH size - 4+	0.214	0.040	0.000
Changed interviewer	-1.508	0.034	0.000
Renter	0.179	0.028	0.000
Rural area	0.504	0.051	0.000



Period 5 Model Propensity Distribution

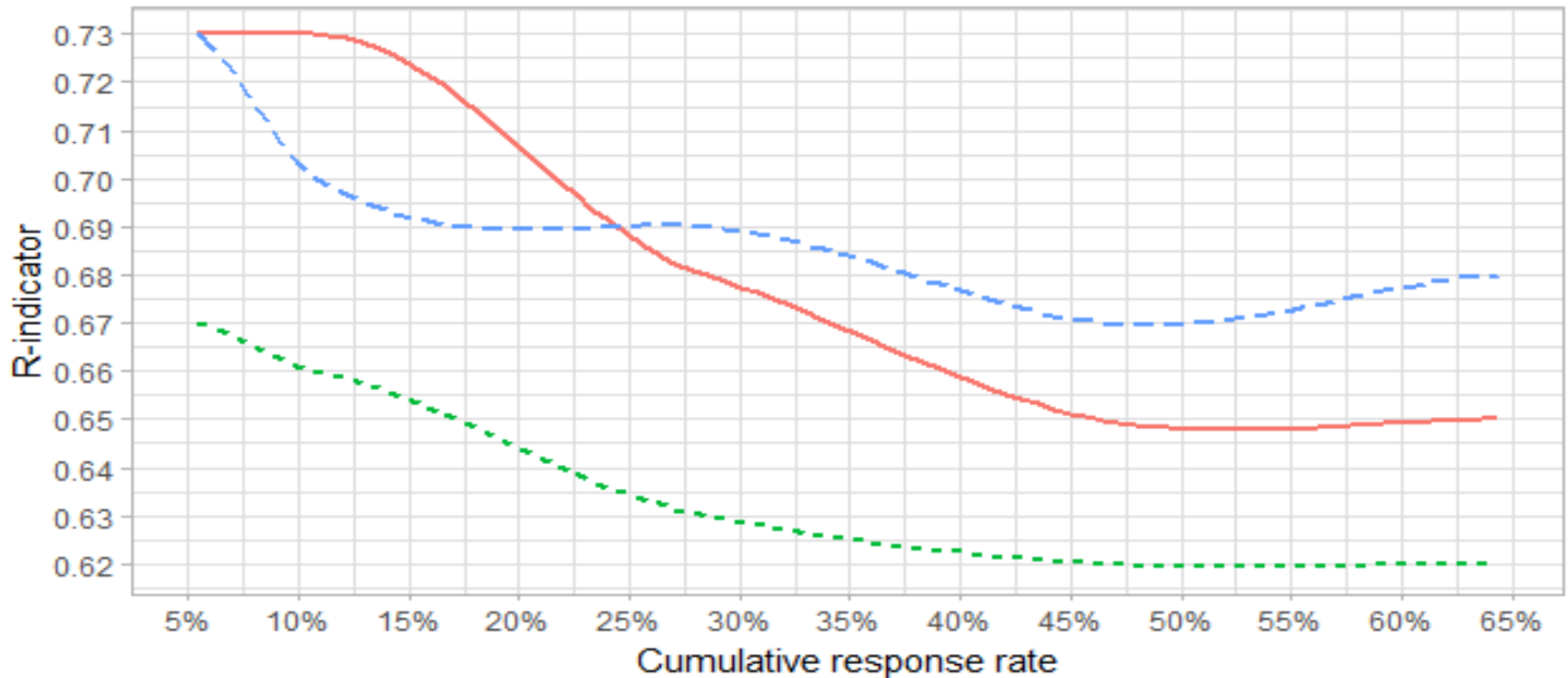


R-Indicators by Period and Model

Period	Months of Data	Classification Tree	Logistic Regression	LASSO Regression
1	1	0.730	0.668	0.734
2	2	0.702	0.658	0.729
3	5	0.690	0.632	0.682
4	8	0.671	0.617	0.653
5	11	0.678	0.622	0.657

R-Indicators

CEQ 2015 Collection Year

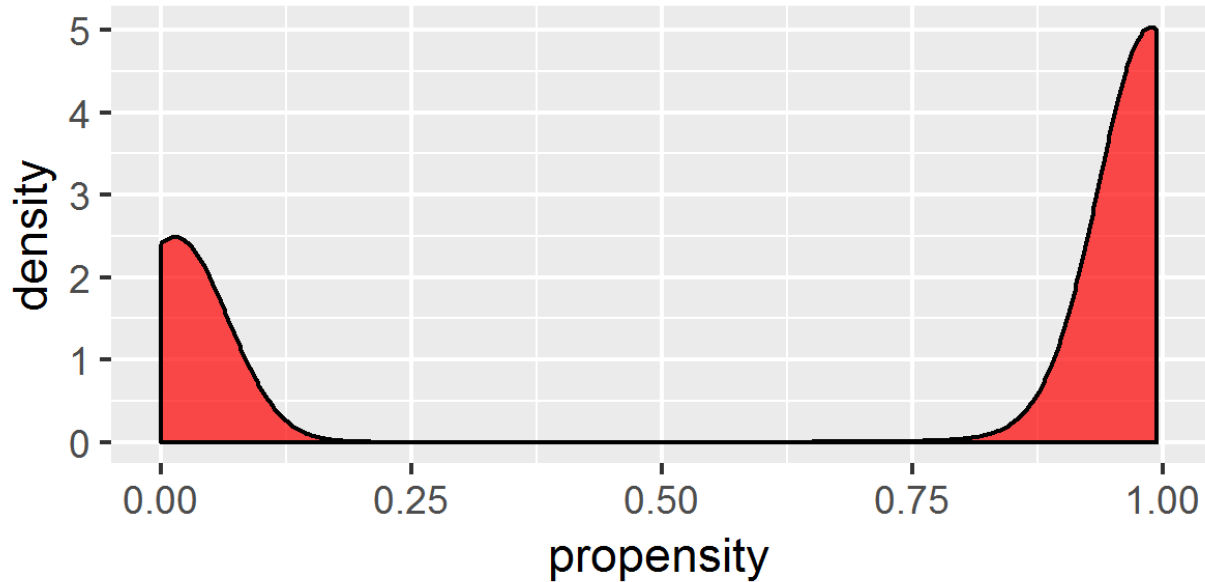


Response propensity model: — lasso_logistic - - - logistic - - - tree



Lessons Learned

Distribution of estimated unit response propensities with variable Language included as a predictor (Accuracy > 98%)



Lessons Learned

- No. sample units with missing value for LANGUAGE: 10,536
 - ▶ No. of survey non-respondents = 10,393
 - ▶ => 98.7% of sample units with missing value for LANGUAGE were non-respondents

KNOW YOUR DATA!!!



Contact Information

Arcenis Rojas

Economist

202-691-6884

rojas.arcenis@bls.gov

Lucilla Tan

Senior Economist

202-691-5128

tan.lucilla@bls.gov

Division of Consumer Expenditure

Surveys

www.bls.gov/cex

