PREDICTING CYBERSECURITY INCIDENTS

Rachel M. Adler

QUESTIONS

1. Does engaging in good cybersecurity habits adequately protect individuals from online attacks?

2. Are certain groups of individuals more vulnerable to such attacks?

THE DATA

- Pew Research Center survey conducted in 2016
- Respondents: 1,040 adult internet users in U.S.
- Three sets of variables
  - 7 cybersecurity incident variables with categorical yes/no responses
  - 20 variables related to cybersecurity habits
  - 16 variables containing demographic information

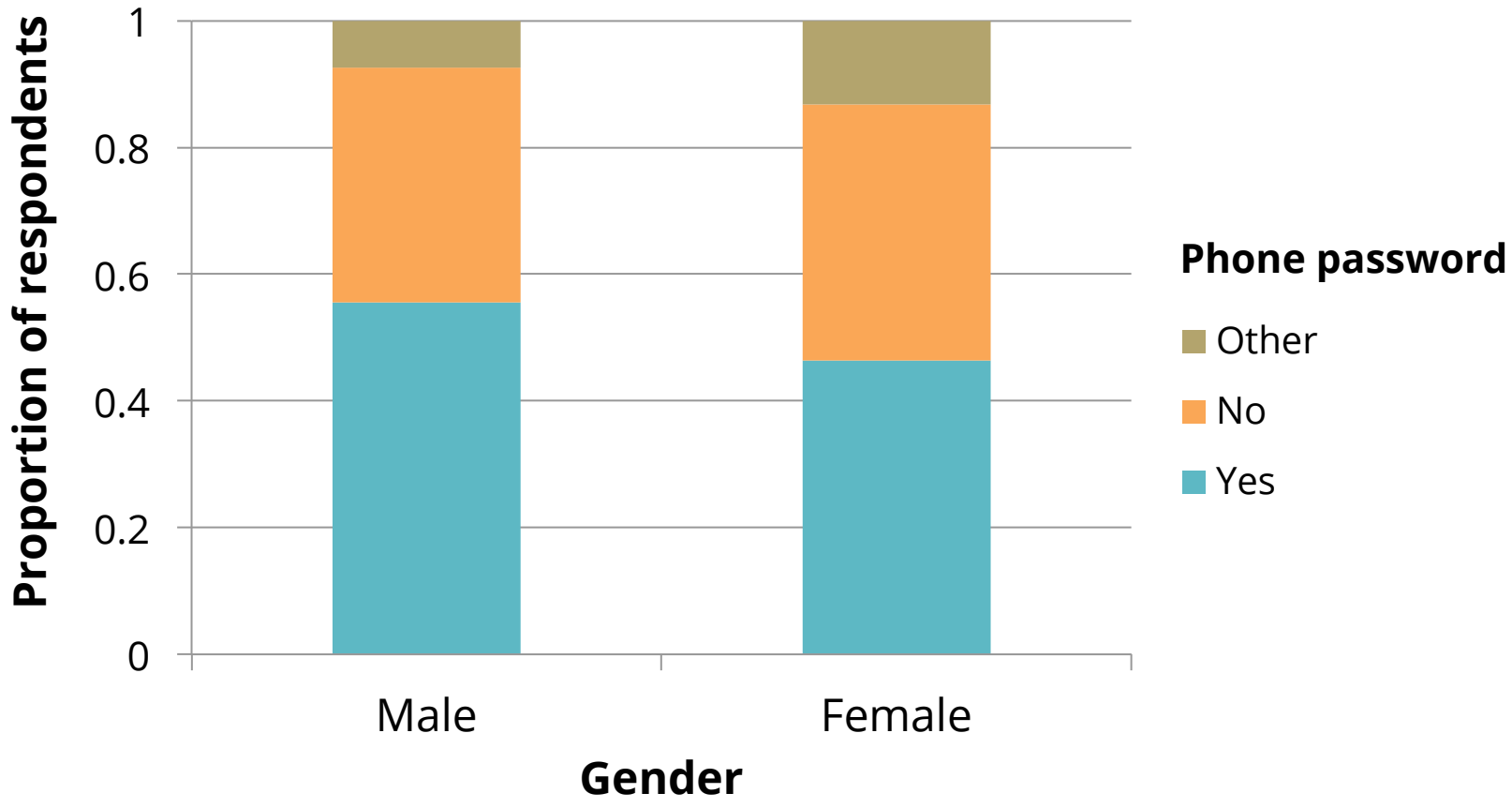
THE DATA

Security incident variables

Name	Description
secur2a	Social security number compromised
secur2b	Other sensitive information compromised
secur2c	Fraudulent charges on credit/debit card
secur2d	Someone took over email account
secur2e	Someone took over social media account
secur2f	Someone opened line of credit/applied for loan in respondent's name
secur2g	Someone received tax refund under respondent's name

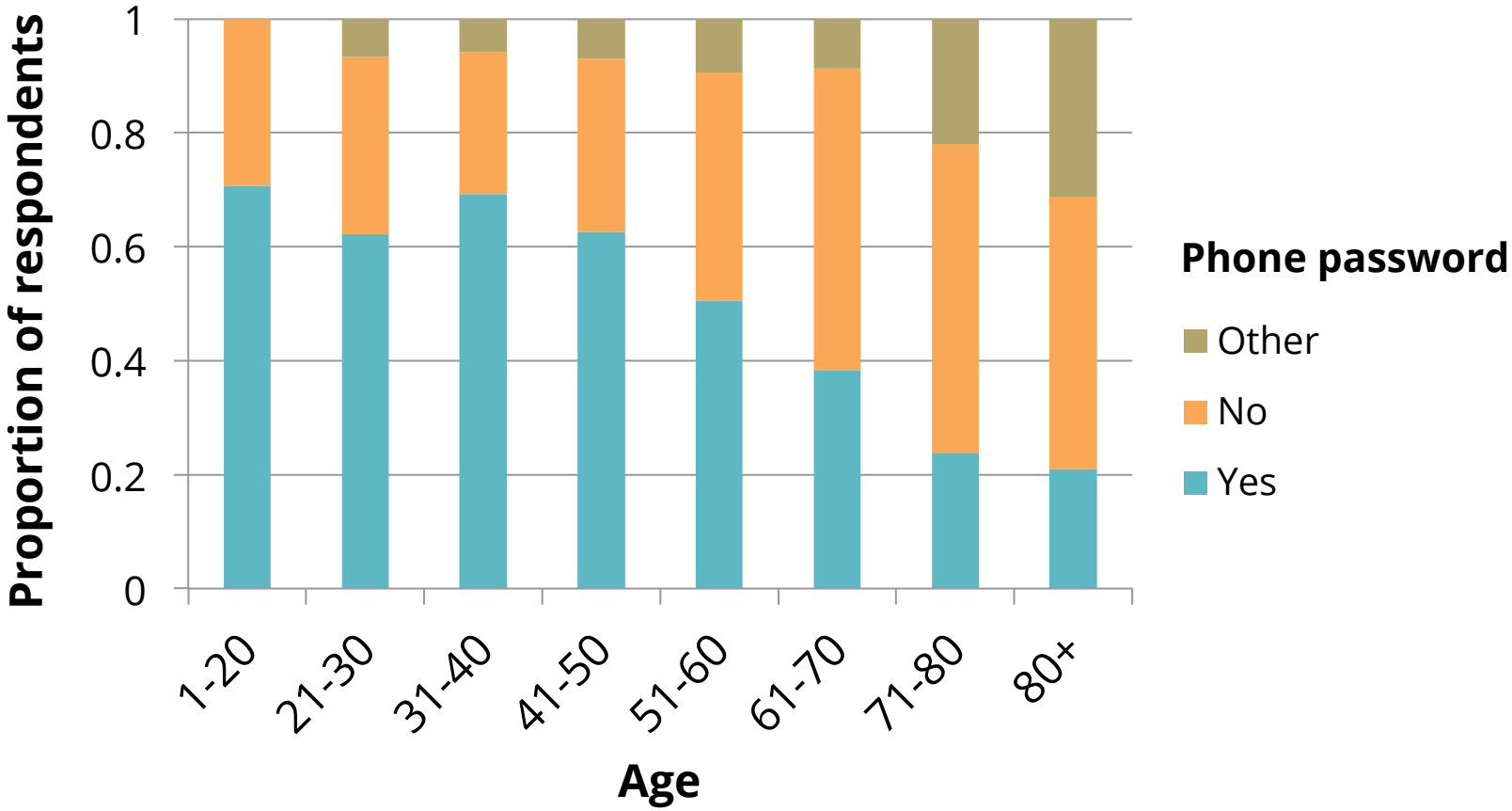
EXPLORATORY DATA ANALYSIS

Gender and use of a phone password



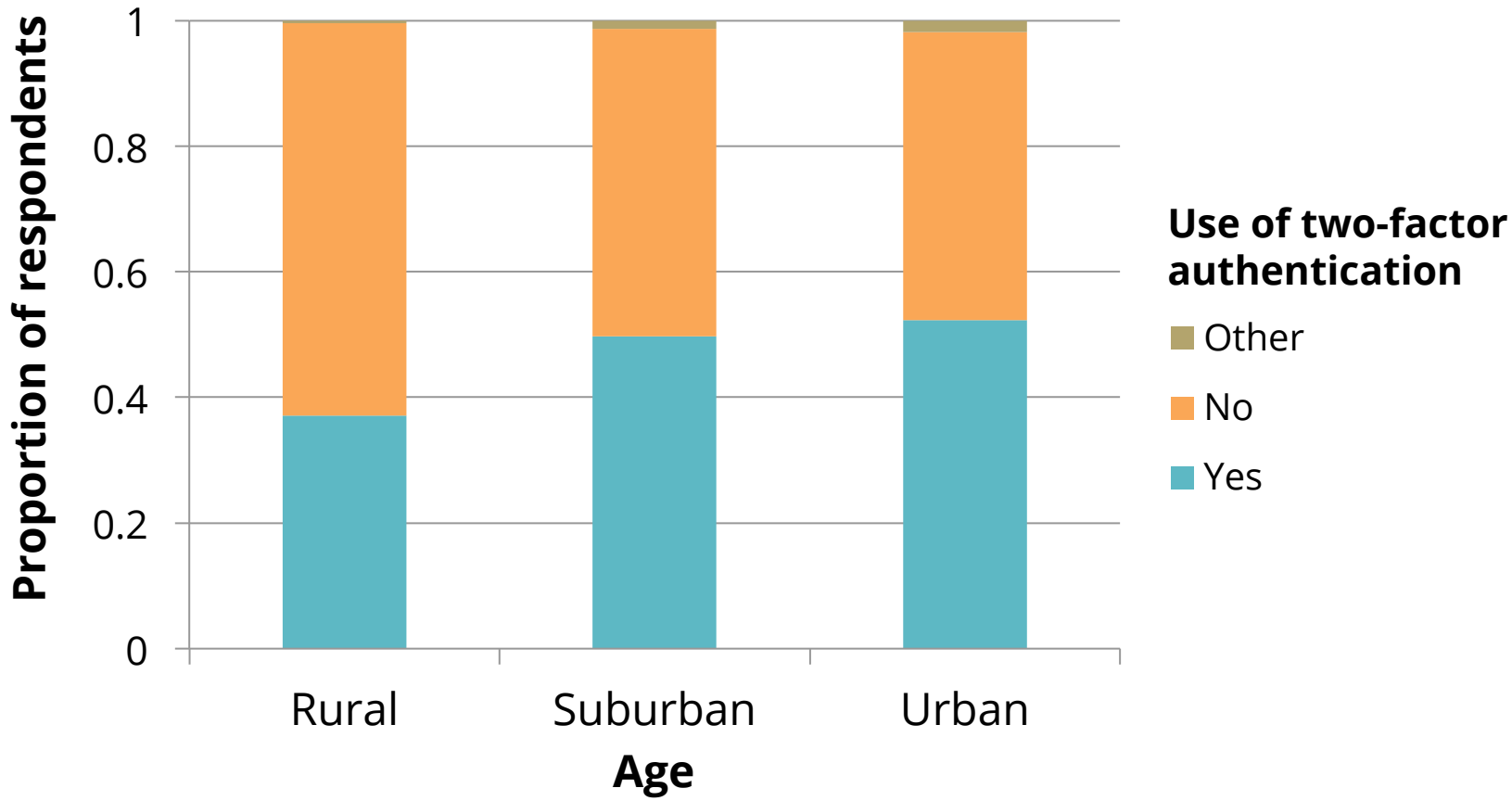
EXPLORATORY DATA ANALYSIS

Age and use of a phone password



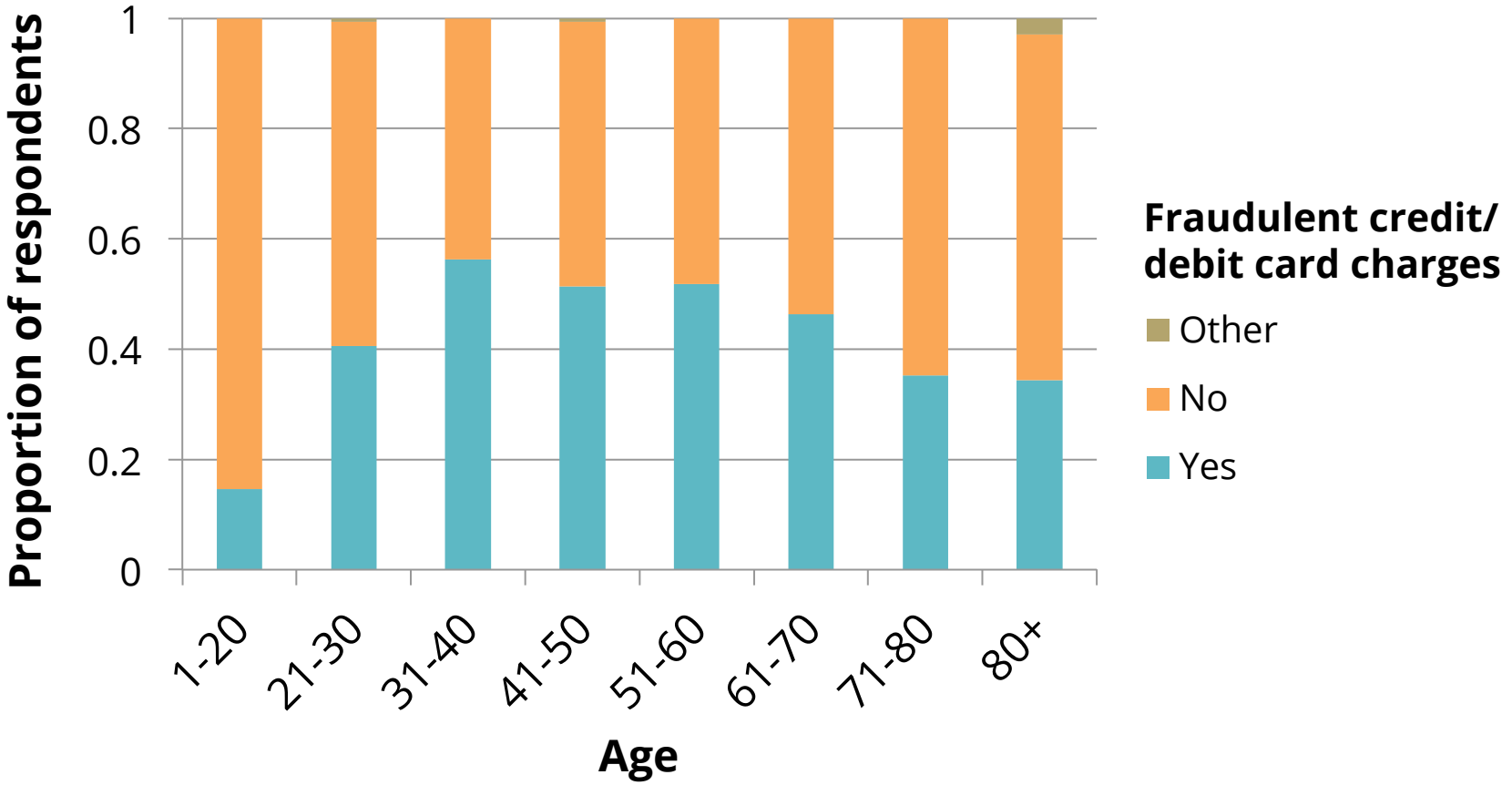
EXPLORATORY DATA ANALYSIS

Region and use of two-factor authentication



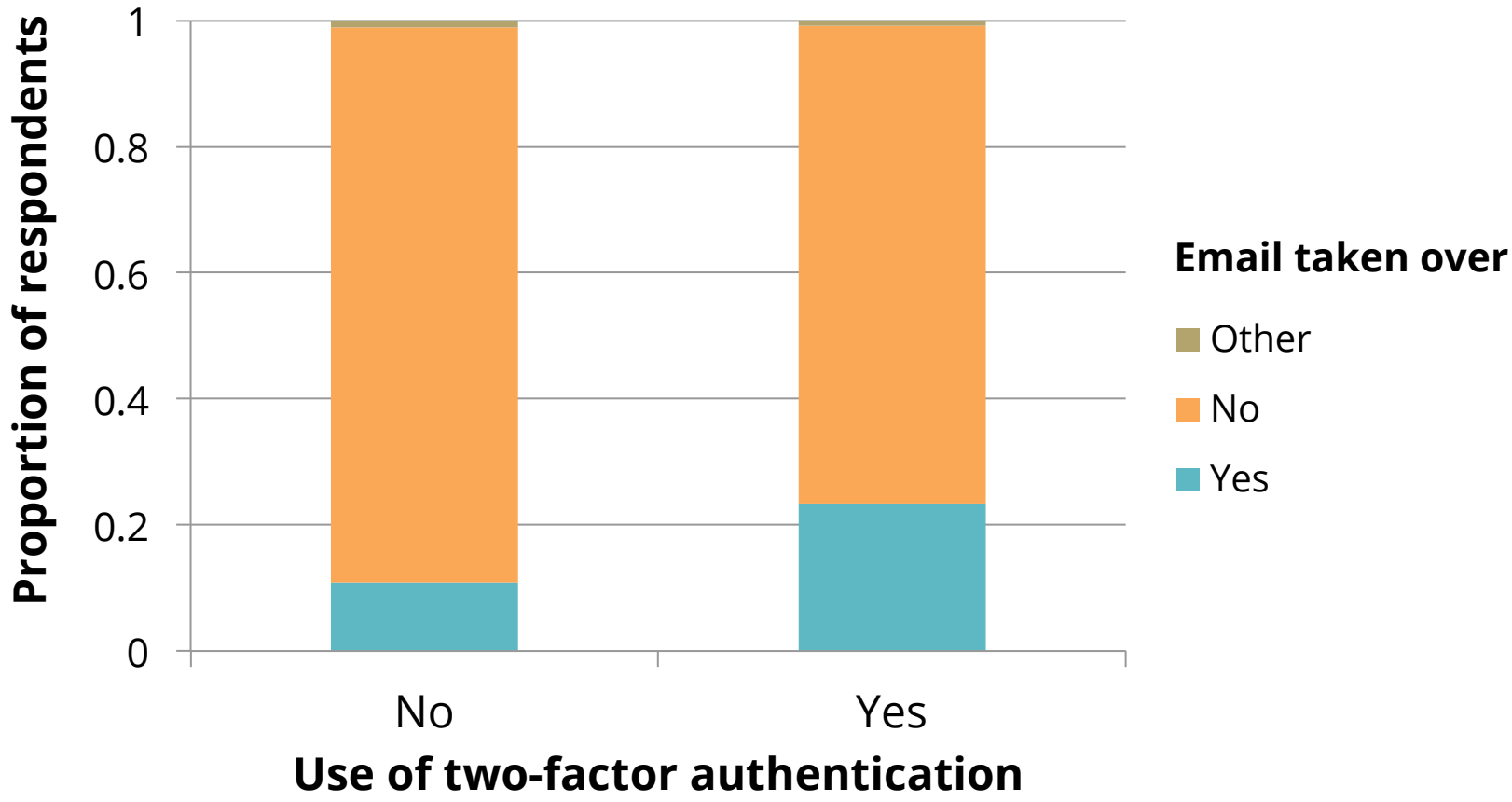
EXPLORATORY DATA ANALYSIS

Age and fraudulent credit/debit card charges



EXPLORATORY DATA ANALYSIS

Use of two factor-authentication and experience of having one's email account taken over



MODELS

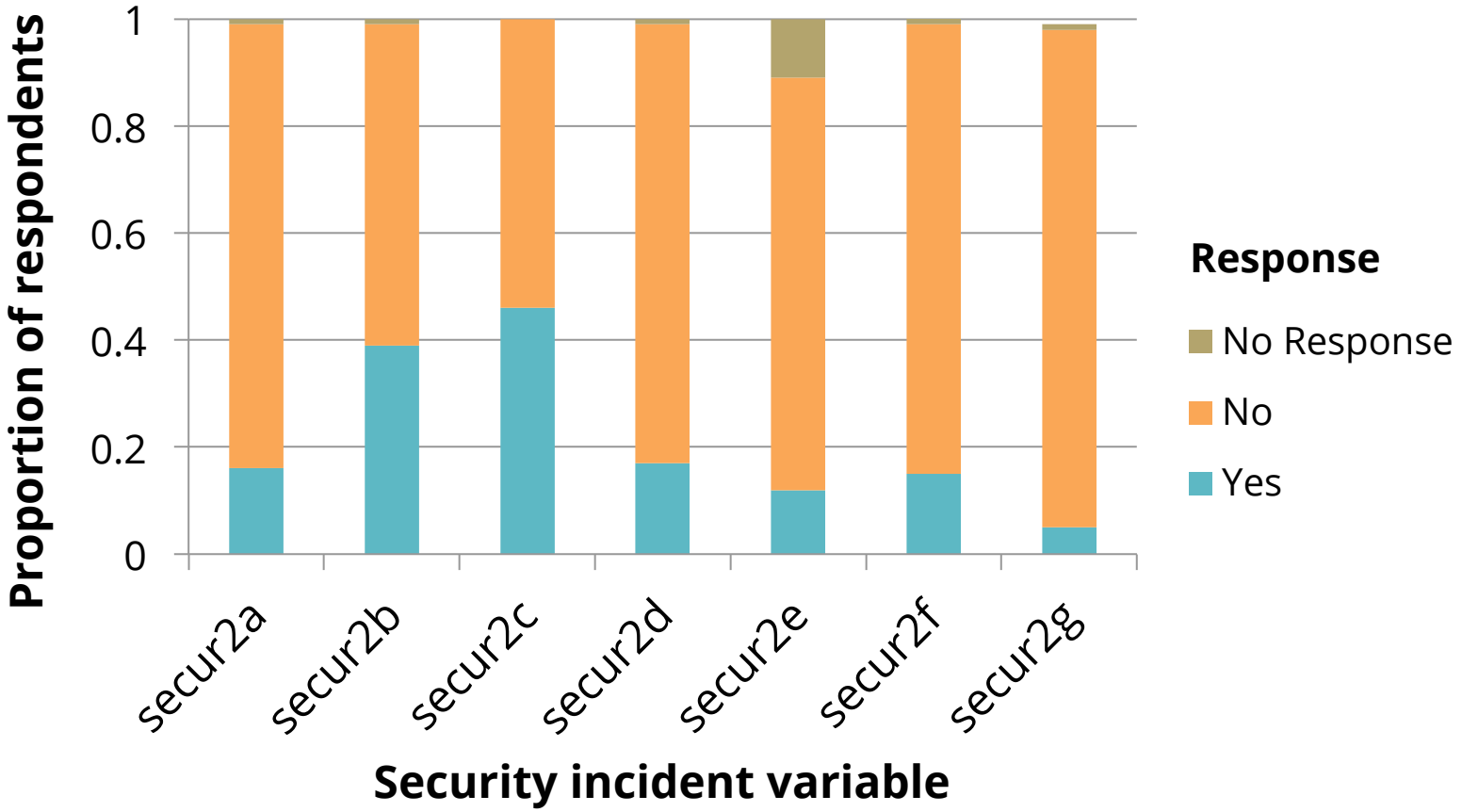
- Models designed to predict each security incident variable from:
  - Demographic variables
  - Security habit variables
- For each security incident variable, generated and compared 4 models
  - Logistic regression
  - k-nearest neighbors
  - Random forest
  - Support vector classifier

MODELS

- Divided data into training and test sets (75-25% split)
- Performed 5-fold cross-validation and identified hyperparameters for each model
  - Used Cohen's kappa as evaluation criterion because it controls for skewed class distributions (see next slide)
- Selected model with highest Cohen's kappa and recall values on test data

MODELS

Distribution of classes for each security incident variable



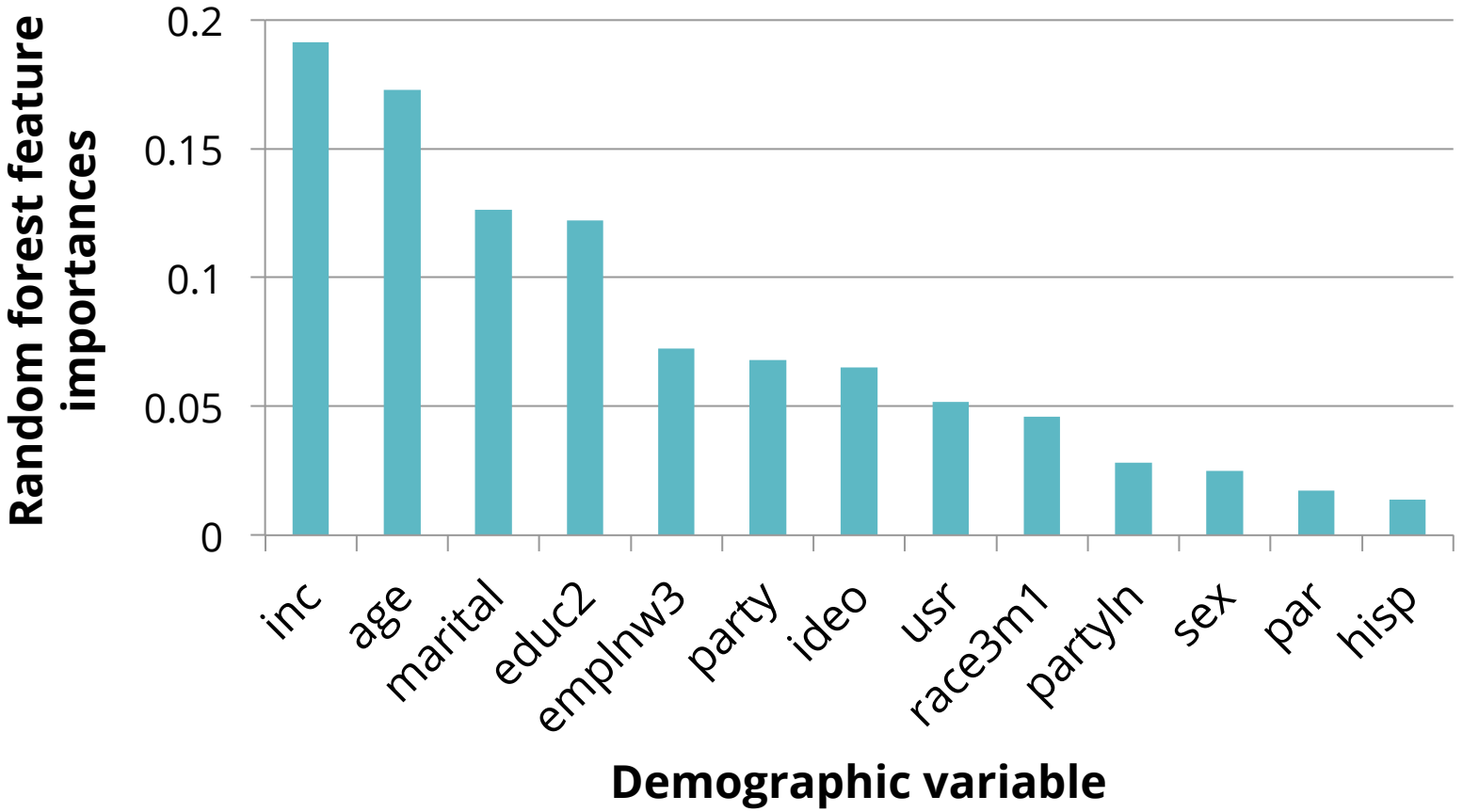
MODELS

Model performance using demographic variables as predictor variables

Name	Model	Cohen's kappa	Recall	Precision	Accuracy	AUC
secur2a	logistic regression	0.08	0.14	0.26	0.80	0.62
secur2b	knn	0.22	0.47	0.55	0.64	0.66
secur2c	random forest	0.27	0.58	0.62	0.64	0.66
secur2d	logistic regression	0.13	0.27	0.28	0.76	0.64
secur2e	logistic regression	0.08	0.23	0.19	0.77	0.60
secur2f	random forest	-0.02	0.13	0.13	0.74	0.47
secur2g	random forest	-0.07	0.00	0.00	0.87	0.52

MODELS

Feature importances for secur2c: fraudulent charges on credit/debit card



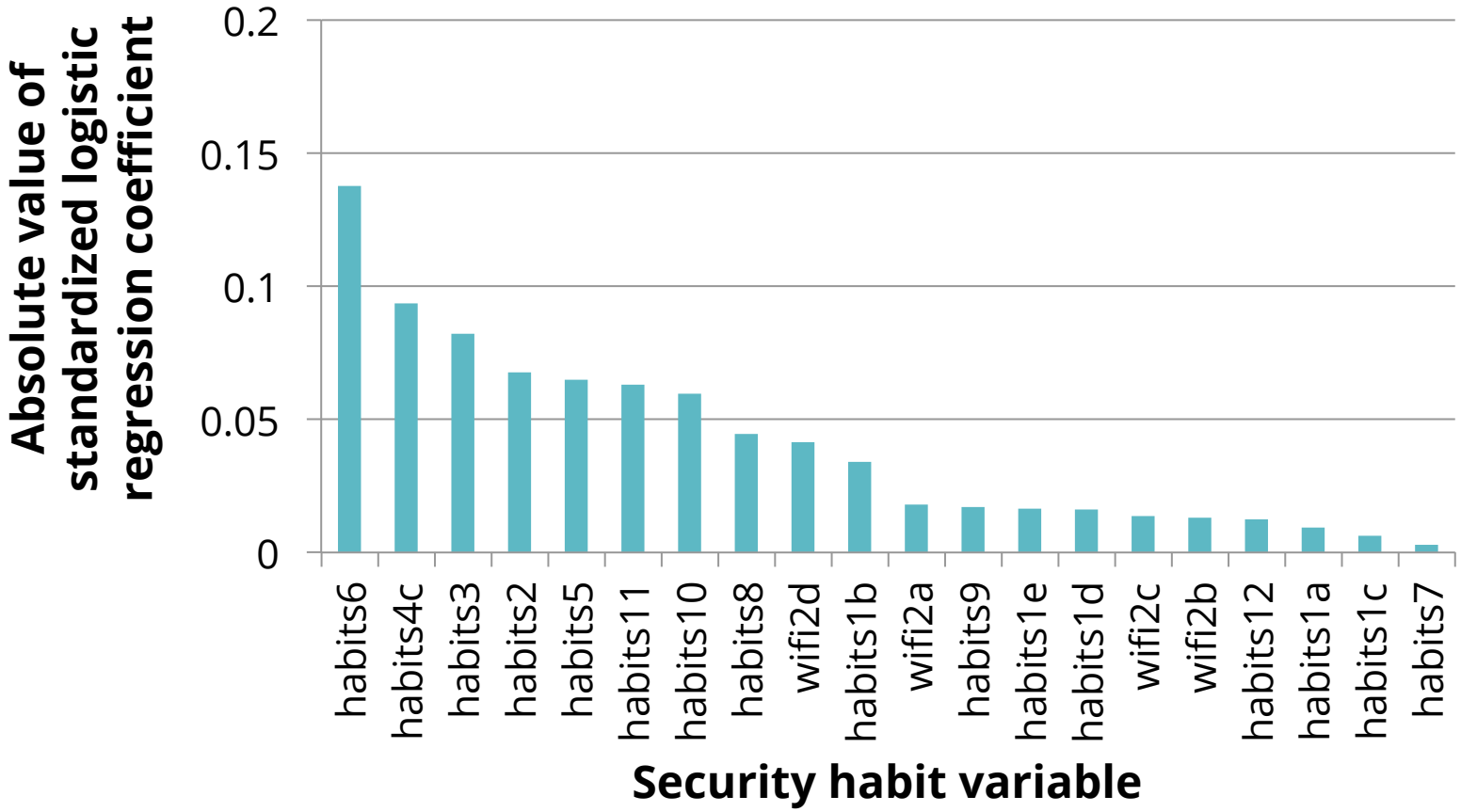
MODELS

Model performance using security habit variables as predictor variables

Name	Model	Cohen's kappa	Recall	Precision	Accuracy	AUC
secur2a	logistic regression	0.06	0.05	0.50	0.84	0.63
secur2b	logistic regression	0.19	0.61	0.49	0.60	0.63
secur2c	logistic regression	0.20	0.87	0.53	0.58	0.64
secur2d	logistic regression	0.15	0.61	0.25	0.62	0.64
secur2e	logistic regression	-0.02	0.13	0.12	0.76	0.61
secur2f	logistic regression	0.08	0.34	0.20	0.70	0.58
secur2g	logistic regression	0.08	0.07	0.20	0.93	0.56

MODELS

Absolute value of standardized coefficients for secur2c: fraudulent charges on credit/debit card



RECOMMENDATIONS

- Generated models may be used to predict cybersecurity incidents from security habit and demographic variables
- Models may also be used to identify particularly vulnerable populations
  - For example: individual's age and income

FUTURE WORK

- Generate model to predict cybersecurity habits from demographic variables
- Add other security habits to survey that might address security incidents models didn't predict well
  - For example: shredding important documents before disposing of them