CAPSTONE PROJECT 1: PREDICTING CYBERSECURITY INCIDENTS

1 PROBLEM DEFINITION

Businesses often promote habits to increase individuals' personal cyber-safety. In order for these campaigns to be effective, it is important to demonstrate a clear link between good cybersecurity habits and actual increased cyber-safety. In addition, it is possible that certain groups of individuals are at increased risk for cyber-security incidents. If so, education efforts should target these particular groups. Thus, this project will show (a) whether engaging in better cybersecurity practices adequately protects individuals from online attacks, and (b) whether certain groups of individuals are more vulnerable to such attacks.

2 CLIENT DESCRIPTION

There are two groups of clients that may profit from this project. The first group of clients comprises organizations that promote cybersecurity, such as the U.S. Department of Homeland Security or the National Integrated Cyber Education Research Center (https://nicerc.org/). These groups seek to educate Americans about good cybersecurity habits. The second group of clients is businesses that strive to increase their customers' cybersecurity efforts to improve their company's overall security. For both types of clients, this project will provide evidence to be shared with consumers that may help encourage better habits. In addition, this project will highlight particular groups of individuals that these organizations should focus their efforts on.

3 DATA OVERVIEW

The data for this project come from a Pew Research Center survey conducted in June 2016. The survey respondents consisted of 1,040 adult internet users in the U.S. All data are publically available via the Pew Research Center (http://www.pewinternet.org/2017/01/26/americans-and-cybersecurity/). The complete dataset contains 1040 rows (observations) and 121 columns (variables). Of these 121 columns/variables, seven are related to security incidents an individual has experienced (see Table 1); all data are categorical (1 = yes, 2 = no). Twenty variables are related to individuals' cyber-security habits (see Appendix). All of these data are categorical; some are binary (1 = yes, 2 = no) and some have more than two responses (e.g., type of security feature used to access phone: pin, password, thumbprint, other). Finally, there are sixteen variables containing demographic information (see Appendix). Some of these data are continuous (e.g., age), some are ordinal (e.g., highest level of education), and some are categorical (e.g., race).

Table 1. Security incident variable names and descriptions

Variable name	Variable 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

4 DATA WRANGLING

As a first step, the relevant columns in the dataset were identified. These columns included data about respondents' (a) cybersecurity habits, (b) cybersecurity incidents experienced, and (c) demographics. The data in these columns largely consisted of 1s ("yes" responses), 2s ("no" responses), 8s ("don't know" responses), and 9s ("refused"). The data wrangling steps performed on these columns are described below.

A. DATA CLEANING

1. Some of the relevant columns contained strings instead of numbers. However, because machine learning algorithms require all values to be numeric, these columns were converted to integers. For example, one column encodes whether the respondent lives in a rural ("R"), suburban ("S"), or urban ("U") region. These values were converted as follows: R = 1, S = 2, and U = 3.
2. The dataset included some numbers as integers, but also some numbers as strings. All values were converted to integers for consistency and to be able to perform machine learning algorithms.
3. For all yes/no items, 2s ("no" responses) were converted to -1s. This means that after replacing blanks with 0s (see below), all yes/no items contain three response values: -1 (no), 0 (blank/other), and 1 (yes). By doing this, the blank/other responses fall in between the "yes" and "no" responses.

B. MISSING VALUES

1. The dataset contained many blank values. All blank values in the relevant columns were replaced with zeros. Zero does not correspond to any other response in these columns.
2. All 8 ("don't know") and 9 ("refused") responses were also replaced with zeros.
3. For some of the missing values, a response could be inferred from another item. When this was the case, the inferred response was entered instead of the missing value. For example, one of the security incident items asks whether the respondent's email account has ever been hacked. However, this question was only asked if the respondent previously indicated they use the internet. If they didn't use the internet, this item was left blank. Importantly, if a respondent doesn't use the internet, then they've also never had their email account hacked, so it makes sense to replace the missing value with a "no" (2) response for this item.

5 EXPLORATORY DATA ANALYSIS

A. DEMOGRAPHIC PATTERNS

Three demographic variables were examined: gender, age, and U.S. region. For each variable, analyses focused on cybersecurity habits and cybersecurity incidents experienced.

i. Gender

Cybersecurity habits. As shown in Figure 1, there was a significant difference between genders in their use of a phone password, $\chi^2(1, N = 932) = 4.02$, $p = .045$. A larger percentage of men (56%) reported using a phone password compared to women (46%). Analyses also revealed a significant difference between genders in their use of two-factor authentication, $\chi^2(1, N = 1026) = 6.14$, $p < .05$. A larger percentage of men (52%) reported using two-factor authentication compared to women (44%).

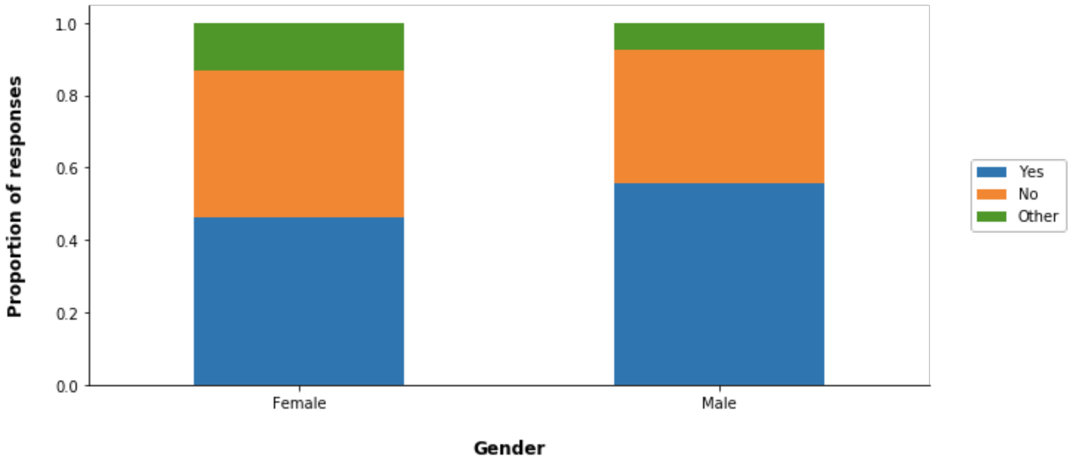

Figure 1. Proportion of each gender that uses a phone password

Cybersecurity incidents. Chi-square tests of independence were conducted to examine the relationship between gender and (a) frauduluent card charges, (b) the likelihood of the respondent's social security number being compromised, and (c) the likelihood of someone opening a line of credit or applying for a loan in the respondent's name. In all cases, there was no significant difference between genders (*ps* > .05).

ii. Age

Respondents were divided into eight groups based on their age (see Table 2). For the below analyses, it is important to keep in mind that the distribution of respondents was not even among these groups.

Table 2. Number of respondents by age group

Age group	Number of

	respondents
Under 21	41
21-30	148
31-40	153
41-50	142
51-60	222
61-70	162
71-80	105
Over 80	67

Cybersecurity habits. As shown in Figure 2, there was a significant difference between age groups' use of a phone password, $\chi^2(7, N = 932) = 81.03$, $p < .0001$. Overall, younger adults were more likely to use a phone password than older adults. Analyses also revealed a significant difference between age groups' use of two-factor authentication, $\chi^2(7, N = 1026) = 84.50$, $p < .0001$. Adults between the ages of 21 and 60 were the most likely to use two-factor authentication.

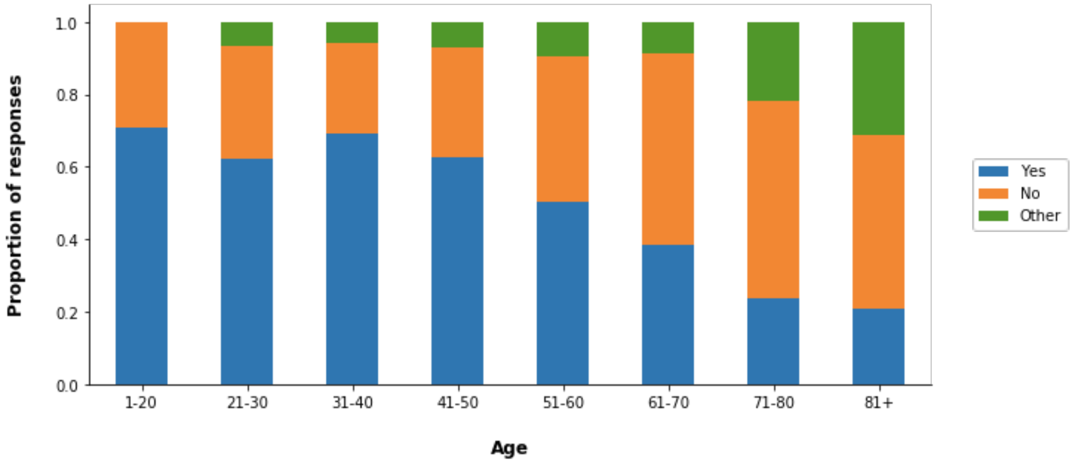

Figure 2. Proportion of each age group that uses a phone password

Cybersecurity incidents. As shown in the figure below, there was a significant difference between age groups' experience of having fraudulent credit/debit chard charges, $\chi^2(7, N= 1036) = 37.01$, $p < .0001$. Adults under 21 were the least likely to have experienced fraudulent credit/debit card charges (15%), while individuals between 31 and 40 were the most likely (56%).

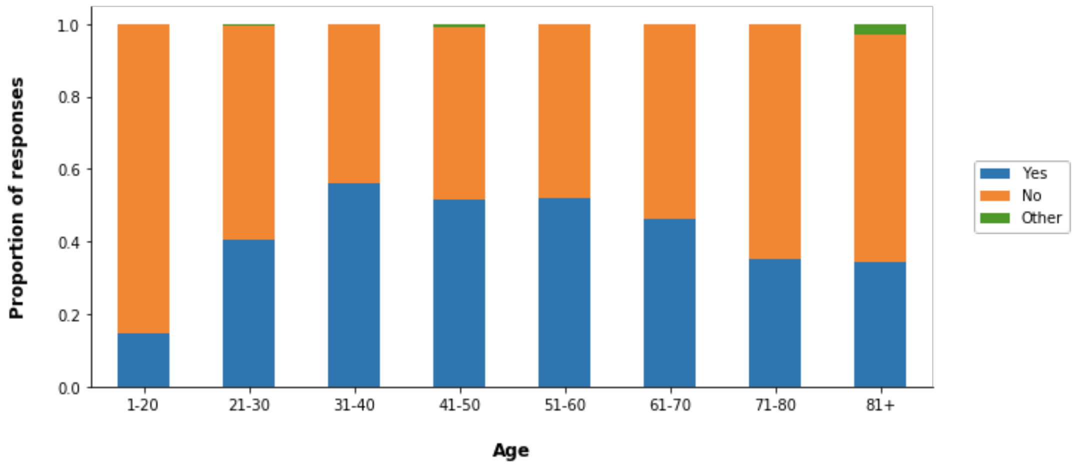

Figure 3. Proportion of each age group that has experienced fraudulent credit/debit card charges

Analyses also revealed a significant difference between age groups' likelihood of having their social security number compromised, $\chi^2(7, N = 1034) = 24.82, p < .0001$. Adults between the ages of 31 and 50 were the most likely to have had their social security number compromised (22%), while adults under 21 were the least likely (2%).

Finally, there was a significant difference between age groups' likelihood of having someone open a line of credit or apply for a loan in the their name, $\chi^2(7, N = 1024) = 14.2, p = .048$. Adults between 51 and 60 were the most likely to have had someone open a line of credit or apply for a loan in their name (19%), while no adults under 21 reported having this experience.

iii. Region

Respondents were divided into three U.S. regions (see Table 3). For the below analyses, it is important to keep in mind that the distribution of respondents was not even among these regions.

Table 3. Number of respondents by region

Region	Number of respondents
Rural	202
Suburban	503
Urban	325

Cybersecurity habits. As shown in Figure 4, there was a significant difference in the frequency of phone password use among individuals from different U.S. regions, $\chi^2(1, N = 932) = 25.51, p < .0001$. While only 34% of respondents in rural regions reported using a phone password, 57% of those in suburban areas and 52% of those in urban areas reported using one.

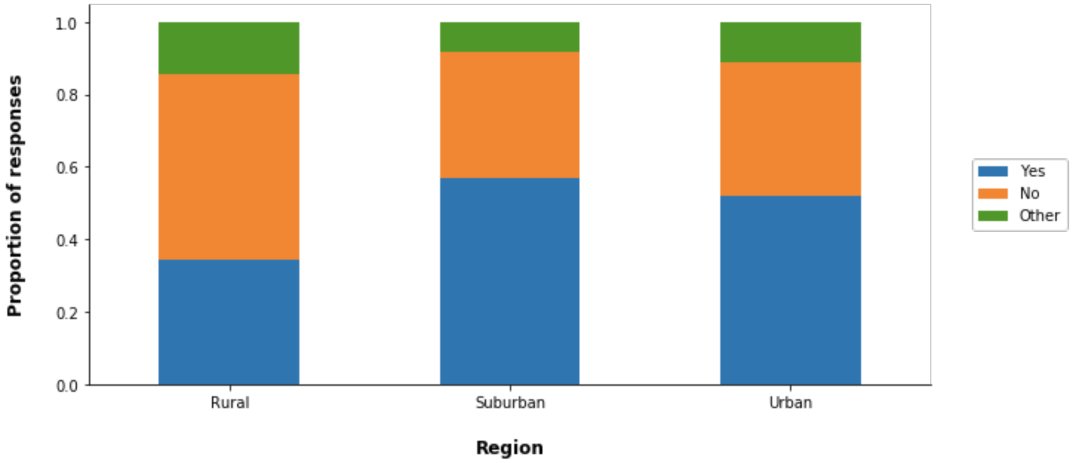

Figure 4. Proportion of each U.S. region that uses a phone password

Analyses also revealed a significant difference in the use of two-factor authentication among individuals from different U.S. regions, $\chi^2(1, N = 1026) = 13.70$, $p = .001$. While only 37% of respondents in rural regions reported using two-factor authentication, 50% of those in suburban areas and 52% of those in urban areas reported using it.

Cybersecurity incidents. As Figure 5 highlights, there was a significant difference between U.S. regions' likelihood of having fraudulent credit/debit chard charges, $\chi^2(2, N= 1036) = 11.59$, $p = .003$. Individuals from rural regions were the least likely to have had fraudulent charges (35%), while individuals from suburban and urban regions were the mostly likely (49% and 48%, respectively).

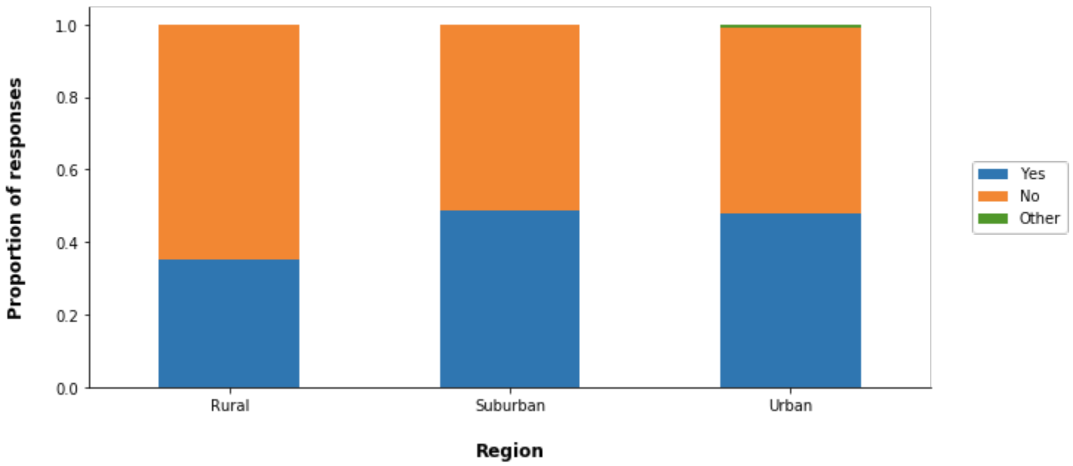

Figure 5. Proportion of each U.S. region that has experienced fraudulent credit/debit card charges

There was no significant difference between regions' experience of having one's social security number compromised ($p > .05$). Similarly, there was no significant difference between regions' experience of having someone open a line of credit or apply for a loan in their name ($p > .05$).

B. CYBER-SECURITY HABITS VS. INCIDENTS

The effect of two sets of cybersecurity habits on the likelihood of experiencing a cybersecurity incident were analyzed: (a) the use of two-factor authentication and (b) the use of public wifi to perform various types of transactions.

iv. Two-factor authentication

As Figure 6 demonstrates, there was a significant relationship between the use of two-factor authentication and the likelihood of having one's email account taken over, $\chi^2(1, N = 1016) = 27.64$, $p < .0001$. Surprisingly, while 23% of individuals who use two-factor authentication had their email account taken over, only 11% of those who don't use two-factor authentication had their email taken over.

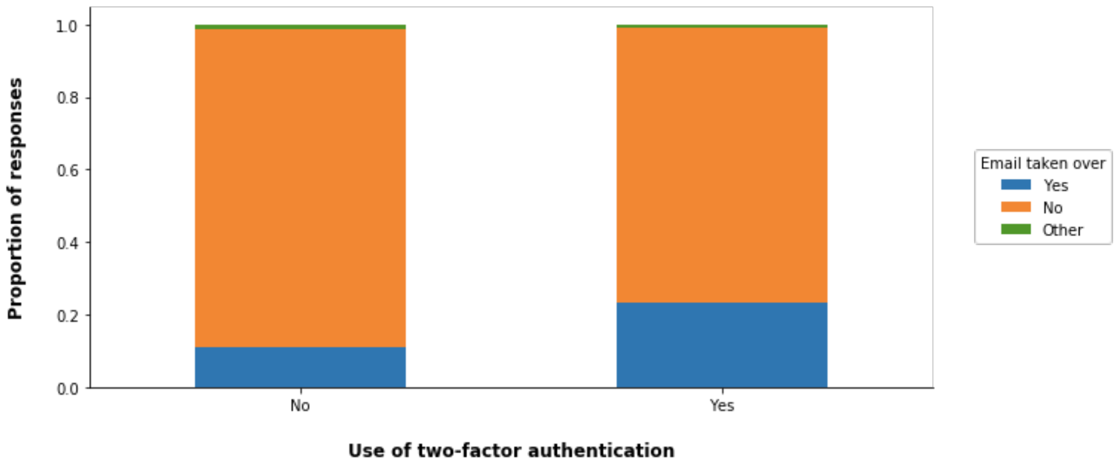

Figure 6. Use of two-factor authentication and the likelihood of having one's email account taken over

Analyses additionally revealed a significant relationship between the use of two-factor authentication and the likelihood of having fraudulent credit/debit card charges, $\chi^2(1, N = 1022) = 43.48$, $p < .0001$ (see Figure 7). Also surprisingly, while 56% of individuals who reported using two-factor authentication had fraudulent credit/debit card charges, only 36% of those who didn't use two-factor authentication had fraudulent charges.

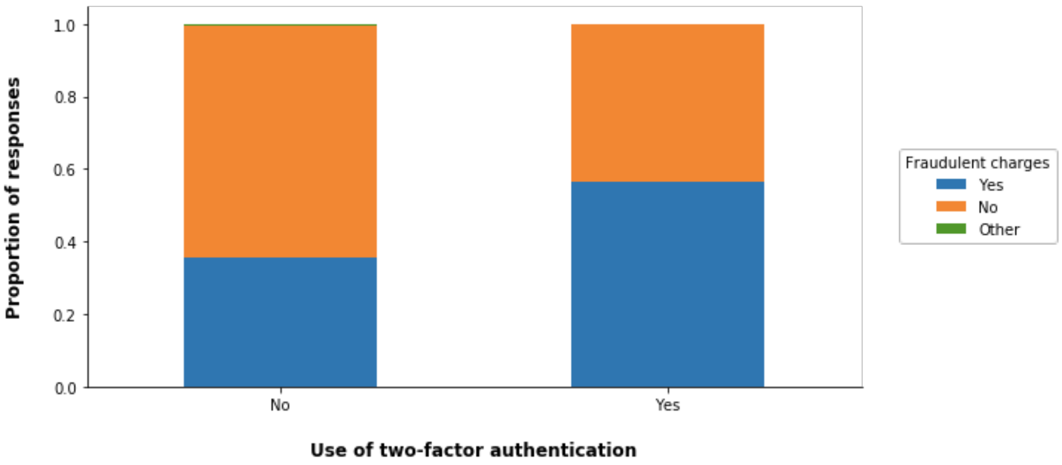

Figure 7. Use of two-factor authentication and the likelihood of having fraudulent credit/debit card charges

There was no statistically significant relationship between the use of two-factor authentication and the likelihood of having someone open a line of credit or apply for a loan in one's name ($p > .05$)

v. Public wifi use

As shown in Figure 8, there was a significant relationship between whether an individual used public wifi to make online purchases, and the likelihood of having fraudulent credit/debit card charges, $\chi^2(1, N = 917) = 9.05$, $p = .003$. While 63% of individuals who made online purchases on public wifi experienced fraudulent charges, only 47% of those who didn't make online purchases had fraudulent charges.

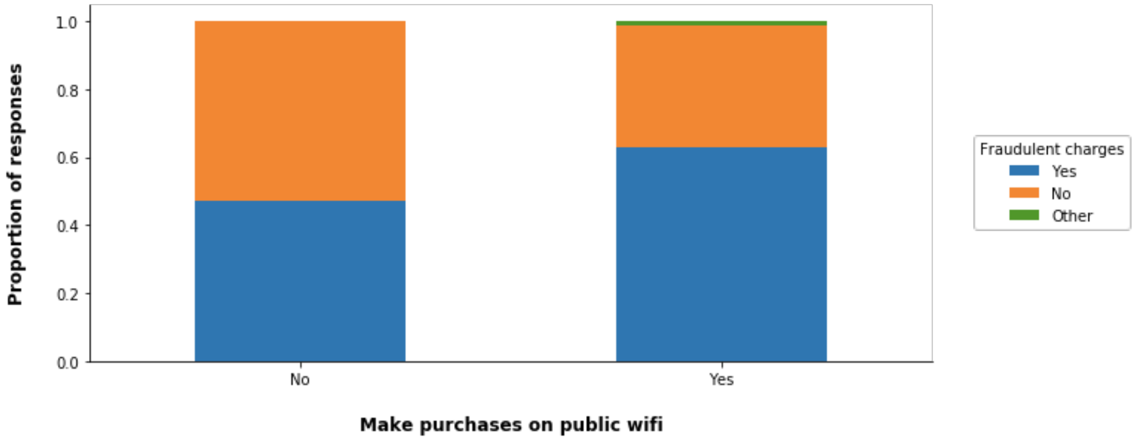

Figure 8. Use of public wifi to make online purchases and the likelihood of having fraudulent credit/debit card charges

There was no statistically significant relationship between whether an individual used public wifi to check their email and the likelihood of their email account being taken over ($p > .05$).

6 MODELING

For each security incident variable, four different machine learning algorithms were developed: random forest, logistic regression, support vector classifier, and k-nearest neighbors. These models were designed to predict the security incident variable from (a) the demographic variables and (b) the security habit variables. For each model, the data were divided into training and test sets (75-25% split) and 5-fold cross-validation was performed to identify the optimal hyperparameters. I selected Cohen's kappa as the evaluation criterion for cross-validation (rather than accuracy), because it controls for skewed class distributions. In this case, there were two classes for each outcome (security incident) variable: "yes" responses and "no" responses. The proportion of each class is depicted in Figure 9. Non-responses ("don't know," "refused," or "NA") were excluded from the models.

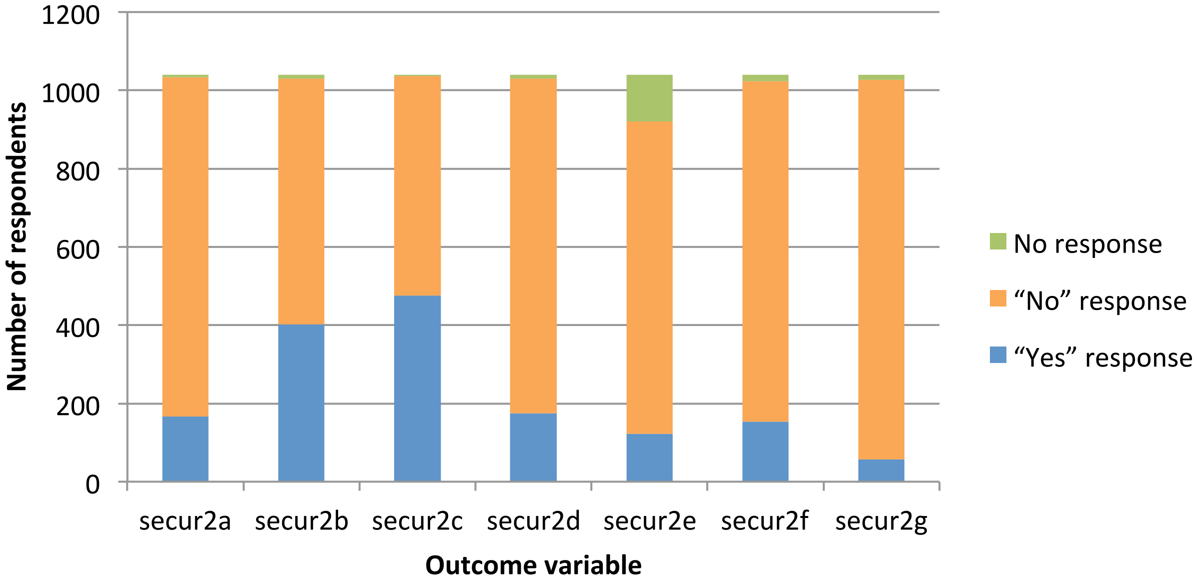

Figure 9. The proportion of each class ("yes", "no", and no response) for each outcome variable

After performing cross-validation and identifying the optimal hyperparameters for each model, the model with the highest Cohen's kappa and recall values on the test data were selected. I selected recall as an additional metric here because it minimizes false negatives, which is particularly important for predicting cybersecurity incidents. The model chosen and the corresponding metrics are depicted in Tables 4 and 5. Interestingly, both sets of models performed particularly well for secur2c (fraudulent charges on credit/debit card), suggesting that the available input variables are especially predictive of this outcome variable.

Table 4. Model performance using demographic variables as predictor variables

Outcome variable	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

Table 5. Model performance using security habit variables as predictor variables

Outcome variable	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

A. DEMOGRAPHIC VARIABLES

As Table 4 depicts, the random forest and logistic regression models were predominantly the best models for the data. Importantly, for some of the variables (e.g., secur2f and secur2g), none of the models performed exceptionally well. Both of these variables had particularly skewed class distributions, and it is therefore possible that this hampered the models' performance. It is also possible that the input variables are simply not predictive of these outcome variables.

For secur2c (fraudulent charges on credit/debit cards), the classifier performed especially well (Cohen's kappa = 0.27). A graph of the feature importances for this model (Figure 10) reveals that income is particularly important. Adults earning below $20,000 were the least likely to have had fraudulent card charges (19%), while adults earning over $50,000 were the most likely (52-60%). Thus, the individuals earning more may be particularly vulnerable. Of course, there are also likely other confounding variables that should be considered before making recommendations. For example, adults earning more may use their credit or debit cards more than adults who earn less, which may lead them to experience more fraudulent charges. Age is also particularly important. As the chi square analyses revealed above, adults between 31 and 40 were the most likely (56%) to have experienced fraudulent credit/debit card charges. Thus, these individuals may also be particularly vulnerable and therefore important for cybersecurity education efforts. Again, however, confounding variables may play a role here. For example, younger adults are less like to *have* credit or debit cards, so they will also be less likely to experience fraudulent charges.

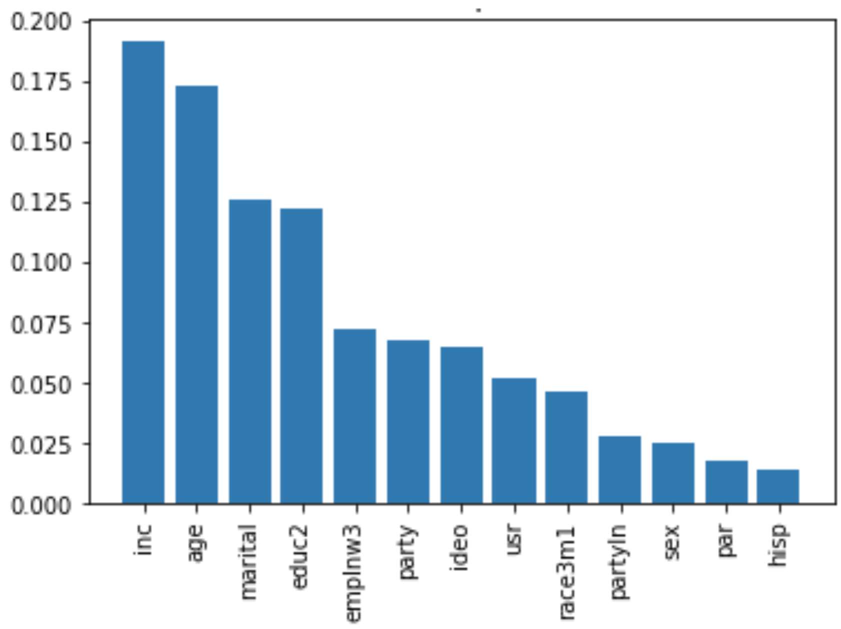

Figure 10. Random forest feature importances for secur2

B. SECURITY HABIT VARIABLES

For security habit variables, logistic regression was largely the best performing model. However, as with the demographic variables, there were some outcome variables for which none of the models performed well (e.g., secur2a, secur2e). Again, both of these variables had particularly skewed class distributions, and it is therefore possible that this hampered the models' performance. It is also possible that the input variables are just not predictive of these outcome variables.

The model performed particularly well for secur2b (other sensitive information compromised) and secur2c (fraudulent charges on credit/debit cards). For secur2c, whether or not the respondent uses two-factor authentication for any online accounts (habits6) was the largest standardized coefficient (see Figure 11). Surprisingly, individuals who do use two-factor authentication are actually *more* likely (56%) to have had fraudulent charges compared to those who don't use two-factor authentication (64%). Again, it's important to consider confounding variables. It may be that individuals who generally have and spend more money are more likely to use two-factor authentication, but they may also be more likely to experience fraudulent charges to begin with. The second most important feature was whether the respondent ever uses passwords that are less secure because more complicated passwords are too hard to remember (habits4c). Here, the pattern is more intuitive: respondents who *do* use less complicated passwords are more likely to have had fraudulent charges (58%) compared to those who *don't* use less complicated passwords (42%).

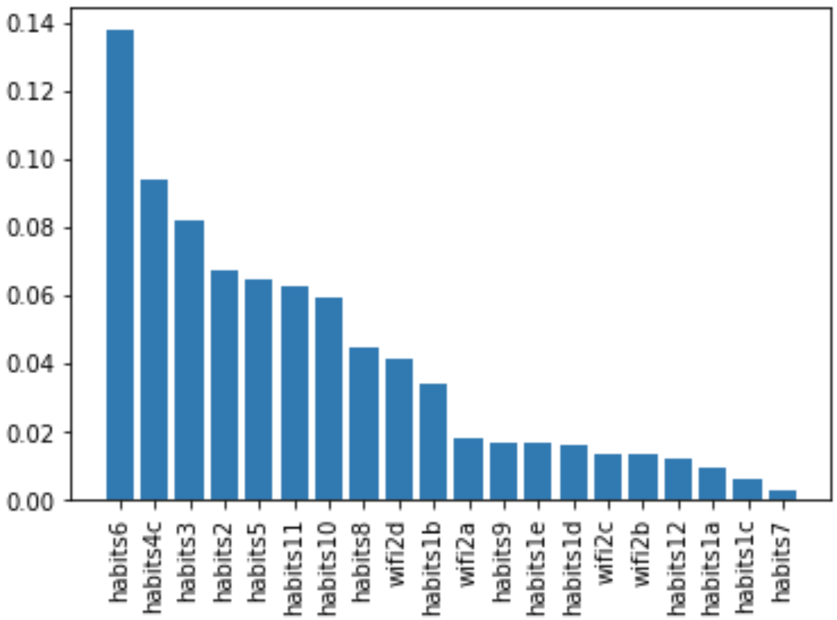

Figure 11. Absolute value of logistic regression standardized coefficients for outcome variable secur2c

7 RECOMMENDATIONS

The models generated in this project may be used to predict certain cybersecurity incident experiences from demographic and cybersecurity habit variables. For most of the security incidents examined here, the models may be used to generate future predictions about individuals' experiences of cybersecurity incidents and identify particularly vulnerable individuals. For example, an individual's age and income may be especially relevant for predicting their experience of cybersecurity incidents. Their cybersecurity habits may also be predictive as well, such as their use of two-factor authentication the types of passwords they use.

8 FUTURE WORK

Future work could generate a model to determine whether a user's cybersecurity habits can be predicted from their demographics. This would further enable educators to target the most relevant populations. In addition, the models may be improved by obtaining more data—particularly from individuals who have experienced some of the less common cybersecurity incidents (e.g., having someone try to receive a tax refund in your name). Furthermore, other security habits could be added to the survey that might address the outcomes that the models didn't predict well. For example, a possible risky habit could be not shredding important documents before disposing of them.

9 APPENDIX

Table 6. Security habit variable names and descriptions

Variable name	Variable description

habits1a	Keep track of passwords by memorizing them
habits1b	Keep track of passwords by writing them on paper
habits1c	Keep track of passwords with password management system
habits1d	Keep track of passwords by saving them in note or document on computer
habits1e	Keep track of passwords by saving them in the internet browser
habits2	Most-used method of keeping track of online passwords
habits3	Passwords for online accounts are mostly similar or different from each other
habits4	Ever use less-secure passwords because complicated ones are hard to remember
habits5	Ever shared password with friend or family member
habits6	Use two-factor authentication for any online accounts
habits7	Ever use social media account information to log into another site
habits8	Use code, password, or other security feature to access phone
habits9	Type of security feature used to access phone
habits10	Frequency/method of updating apps on phone
habits11	Frequency/method of updating OS on phone
habits12	Virus protection apps on phone
wifi2a	While on public wifi, ever make online purchases
wifi2b	While on public wifi, ever do online banking transactions
wifi2c	While on public wifi, ever use social media
wifi2d	While on public wifi, ever use email

Table 7. Demographic variable names and descriptions

Variable name	Variable description
sex	Sex
age	Age
educ2	Highest level of school completed/degree received
hisp	Of Hispanic, Latino, or Spanish origin
race3m1	Race
marital	Marital status
par	Parent or guardian of any child under 18 now living in house
emplnw3	Employment status
party	Political party
partyln	Political party leaning
ideo	Political views
inc	Total family income before taxes in 2015
usr	Region in U.S.