Predicting Past Year Identity Theft with the 2016 Identity Theft Supplement of the National Crime Victimization Survey

Erika Harrell

12/21/2020

Executive Summary

Data analysis was conducted to see if certain predictor variables were associated with past year identity theft. The data used in this analysis came from the 2016 Identity Theft Supplement (ITS) to the National Crime Victimization Survey (NCVS). The results of this analysis found that identity theft was dependent on the majority of the predictors used (age, race/Hispanic origin, annual household income, being a victim of identity theft prior to the past year, having personal information exposed due to a data breach, and using preventative behaviors). Gender was found to be independent of past year identity theft.A logistic regression model, with all predictors excluding sex, predicting past year identity theft was trained on the data. All predictors were found to be significant in predicting the outcome.

```
#adding libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggthemes)
library(caret)
```

```
## Loading required package: lattice
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(ROCR)
library(Metrics)
```

```
##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:Metrics':
##
##     auc

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Loading Data

Data were downloaded from the 2016 ITS study page from the National Archives of Criminal Justice Data to a subfolder entitled "data" in the project on R-Studio, unzipped, loaded into R-Studio and renamed with a shorter name. Due to the size of the datafile, the memory limit had to be increased prior to loading the data in R-Studio.

```
## [1] 20000
```

Data Wrangling

There were 125,165 total persons in the 2016 ITS. Only completed telephone and personal interviews were used in the analysis which left 96,130 interviews or observations in the dataset.

```
##                          Number
## (1) Personal interview   57119
## (2) Telephone interview  39011
## (5) ITS Noninterview     29035
## Total                   125165
```

```
##                          Number
## (1) Personal interview   57119
## (2) Telephone interview  39011
## (5) ITS Noninterview         0
## Total                    96130
```

Created variables that would be used in analysis from the larger ITS dataset.

Individual variables creatd in the previous step were combined into a smaller dataset and larger dataset was removed.

Summary of the created dataset.

```
## 'data.frame':    96130 obs. of  22 variables:
##  $ sex                  : Factor w/ 3 levels "(1) Male","(2) Female",..: 2 1 1 1 1 2 1 1 2 2 ...
##  $ race                 : Factor w/ 21 levels "(01) White only",..: 1 2 6 1 1 1 1 1 1 1 ...
##  $ hispanic             : Factor w/ 3 levels "(1) Yes","(2) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ income               : Factor w/ 14 levels "(01) Less than $5,000",..: 14 14 14 14 4 13 13 13 13 1
##  $ age                  : num  46 50 22 78 50 30 29 62 60 74 ...
##  $ pastyearbankacct     : Factor w/ 5 levels "(01) Yes","(02) No",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ existing_bank        : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ currentccacct        : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 1 1 1 1 2 1 1 ...
##  $ pastyearccacct       : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 NA NA NA NA 2 NA NA ...
##  $ existing_credit_card: Factor w/ 5 levels "(01) Yes","(02) No",..: NA NA NA 2 2 2 2 NA 2 2 ...
##  $ other_existing_accts: Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ open_new_acct        : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ personal_info        : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ OUTSIDE_PAST_YEAR    : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 1 2 2 1 2 2 2 2 ...
##  $ CHCKD_CR_PAST_YR     : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 1 2 2 1 1 2 2 1 ...
##  $ CHNG_PASSWORDS       : Factor w/ 5 levels "(01) Yes","(02) No",..: 1 1 2 2 2 1 1 2 2 2 ...
##  $ PURCHASE_IDTHFT_INS  : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ SHRED_DOCS           : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 1 1 1 1 1 1 1 2 ...
##  $ VERIFY_CHARGES       : Factor w/ 5 levels "(01) Yes","(02) No",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ PROTECT_COMPUTER     : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ PURCHASE_IDTHFT_PROT: Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ notify_breach        : Factor w/ 5 levels "(01) Yes","(02) No",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
##           sex                                  race              hispanic
##  (1) Male   :44908   (01) White only          :79770   (1) Yes    :12131
##  (2) Female :51222   (02) Black only          :10051   (2) No     :83999
##  (8) Residue:    0   (04) Asian only          : 4160   (8) Residue:    0
##                      (03) Am Ind/AK native only:  656
##                      (07) White-Amer Ind      :  530
##                      (06) White-Black         :  304
##                      (Other)                  :  659
##                      income              age              pastyearbankacct
##  (14) $75,000 and over  :33662   Min.   :16.00   (01) Yes       :85793
##  (13) $50,000 to $74,999:17342   1st Qu.:34.00   (02) No        :10337
##  (12) $40,000 to $49,999: 9330   Median :50.00   (08) Residue   :    0
##  (10) $30,000 to $34,999: 5811   Mean   :49.26   (98) Refused   :    0
##  (11) $35,000 to $39,999: 5316   3rd Qu.:63.00   (99) Don't know:    0
##  (08) $20,000 to $24,999: 5137   Max.   :90.00
##  (Other)                :19532
```

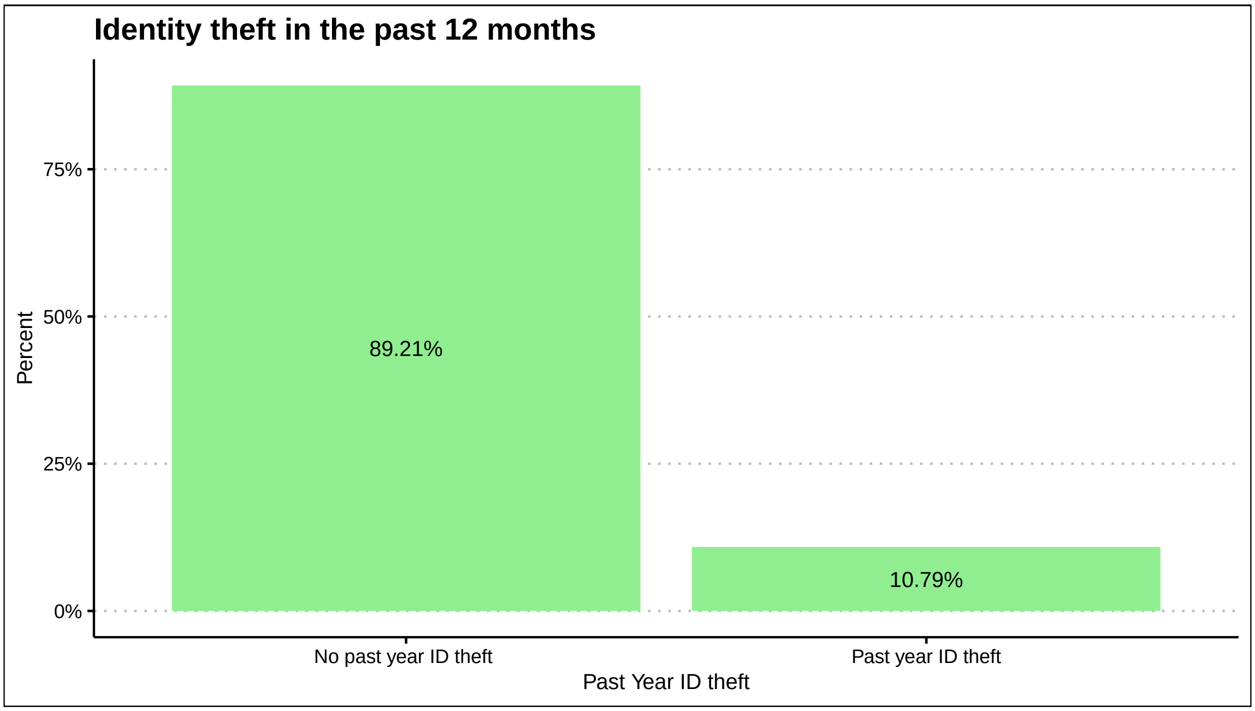
```
##         existing_bank          currentccacct          pastyearccacct
##  (01) Yes       : 4665   (01) Yes       :69446   (01) Yes       : 1003
##  (02) No        :81128   (02) No        :26670   (02) No        :25681
##  (08) Residue   :    0   (08) Residue   :    0   (08) Residue   :    0
##  (98) Refused   :    0   (98) Refused   :   10   (98) Refused   :    0
##  (99) Don't know:    0   (99) Don't know:    4   (99) Don't know:    0
##  NA's           :10337                           NA's           :69446
##
##      existing_credit_card      other_existing_accts        open_new_acct
##  (01) Yes       : 5460     (01) Yes       :  815     (01) Yes       :  589
##  (02) No        :64989     (02) No        :95315     (02) No        :95541
##  (08) Residue   :    0     (08) Residue   :    0     (08) Residue   :    0
##  (98) Refused   :    0     (98) Refused   :    0     (98) Refused   :    0
##  (99) Don't know:    0     (99) Don't know:    0     (99) Don't know:    0
##  NA's           :25681
##
##         personal_info         OUTSIDE_PAST_YEAR       CHCKD_CR_PAST_YR
##  (01) Yes       :  473   (01) Yes       :12267   (01) Yes       :44385
##  (02) No        :95657   (02) No        :83692   (02) No        :51344
##  (08) Residue   :    0   (08) Residue   :   48   (08) Residue   :   80
##  (98) Refused   :    0   (98) Refused   :   39   (98) Refused   :  158
##  (99) Don't know:    0   (99) Don't know:   84   (99) Don't know:  163
##
##
##         CHNG_PASSWORDS       PURCHASE_IDTHFT_INS          SHRED_DOCS
##  (01) Yes       :36861   (01) Yes       :12149    (01) Yes       :67772
##  (02) No        :58670   (02) No        :83440    (02) No        :27942
##  (08) Residue   :   88   (08) Residue   :   88    (08) Residue   :   92
##  (98) Refused   :  250   (98) Refused   :  193    (98) Refused   :  196
##  (99) Don't know:  261   (99) Don't know:  260    (99) Don't know:  128
##
##
##         VERIFY_CHARGES         PROTECT_COMPUTER      PURCHASE_IDTHFT_PROT
##  (01) Yes       :75419   (01) Yes       :16447   (01) Yes       : 4881
##  (02) No        :20344   (02) No        :79001   (02) No        :90756
##  (08) Residue   :   94   (08) Residue   :  100   (08) Residue   :  100
##  (98) Refused   :  180   (98) Refused   :  209   (98) Refused   :  217
##  (99) Don't know:   93   (99) Don't know:  373   (99) Don't know:  176
##
##
##         notify_breach
##  (01) Yes       :11037
##  (02) No        :84652
##  (08) Residue   :  102
##  (98) Refused   :  179
##  (99) Don't know:  160
##
##
```

Recoded variables, collapsing categories and computing necessary categories and variables for analysis.

Exploratory analysis

Past year identity theft

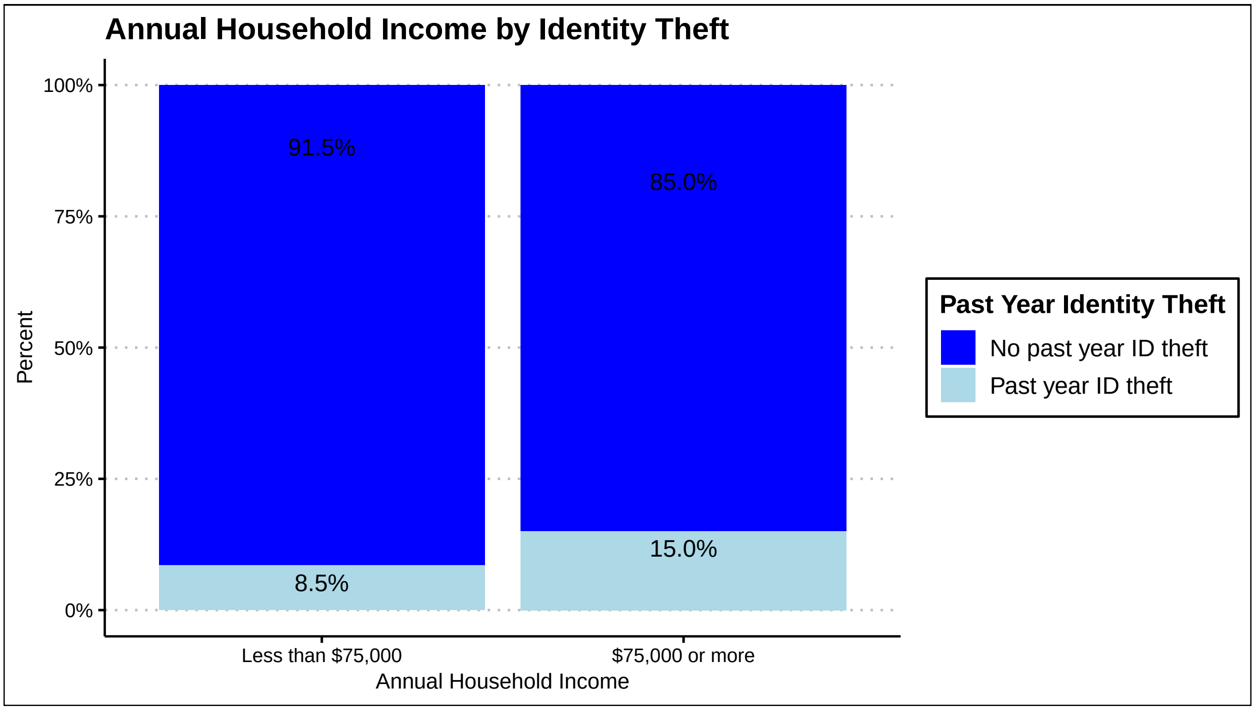
About 11% of the sample reported at least one type of identity theft (misuse of an existing account, misuse of personal information to open new account, or misuse of personal information for other fraudulent purposes) in the past year while 89% of the sample reported no identity theft.



Annual household income

About two thirds of the sample were in households with annual incomes of less than $75,000 (65%) while the remainder (35%) were in households with annual incomes of at least $75,000. Within each income category the majority of respondents did not report experiencing identity theft in the past year. However, 15% of persons in households with incomes of $75,000 or more reported past year identity theft, compared to 9% of those in other households.

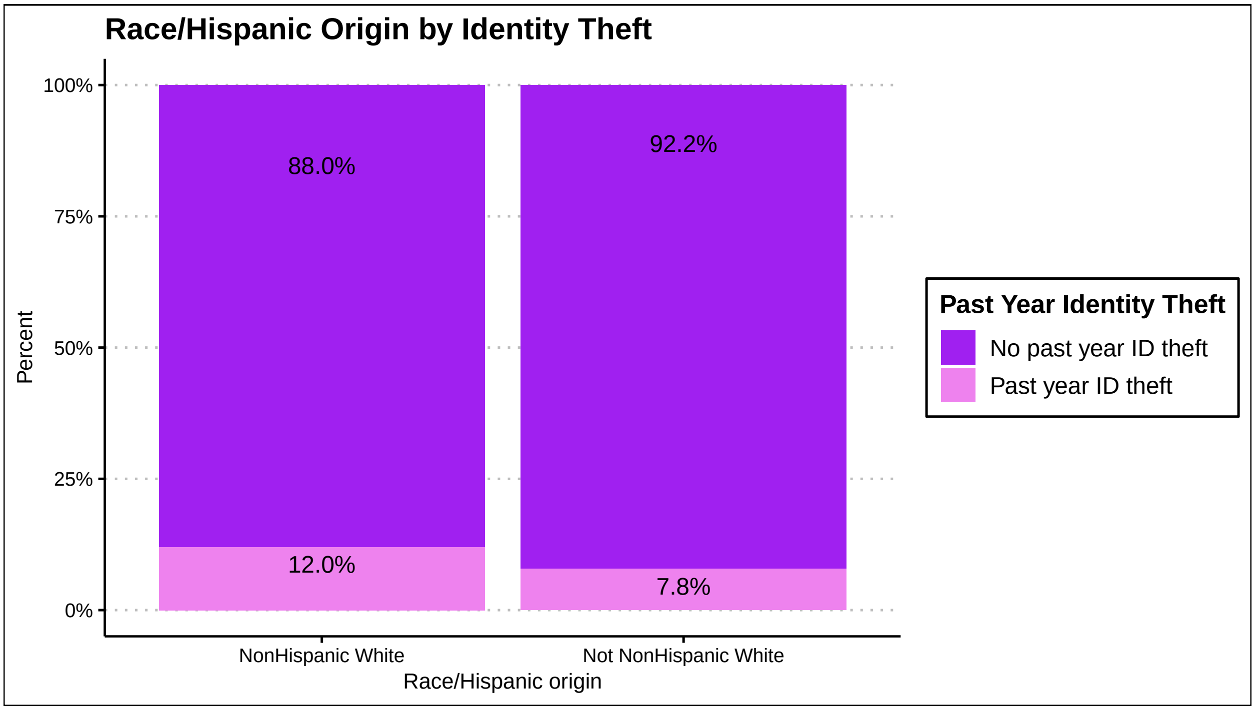
```
##                   Count Percentage
## Total             96130        100
## Less than $75,000 62468         65
## $75,000 or more   33662         35
```



Race/Hispanic origin

Seventy one percent of respondents were NonHispanic White while 29% were belonged to other race/Hispanic origin groups. Twelve percent of nonHispanic Whites reported past year identity theft compared to 8% of other persons reported identity theft in the past year.

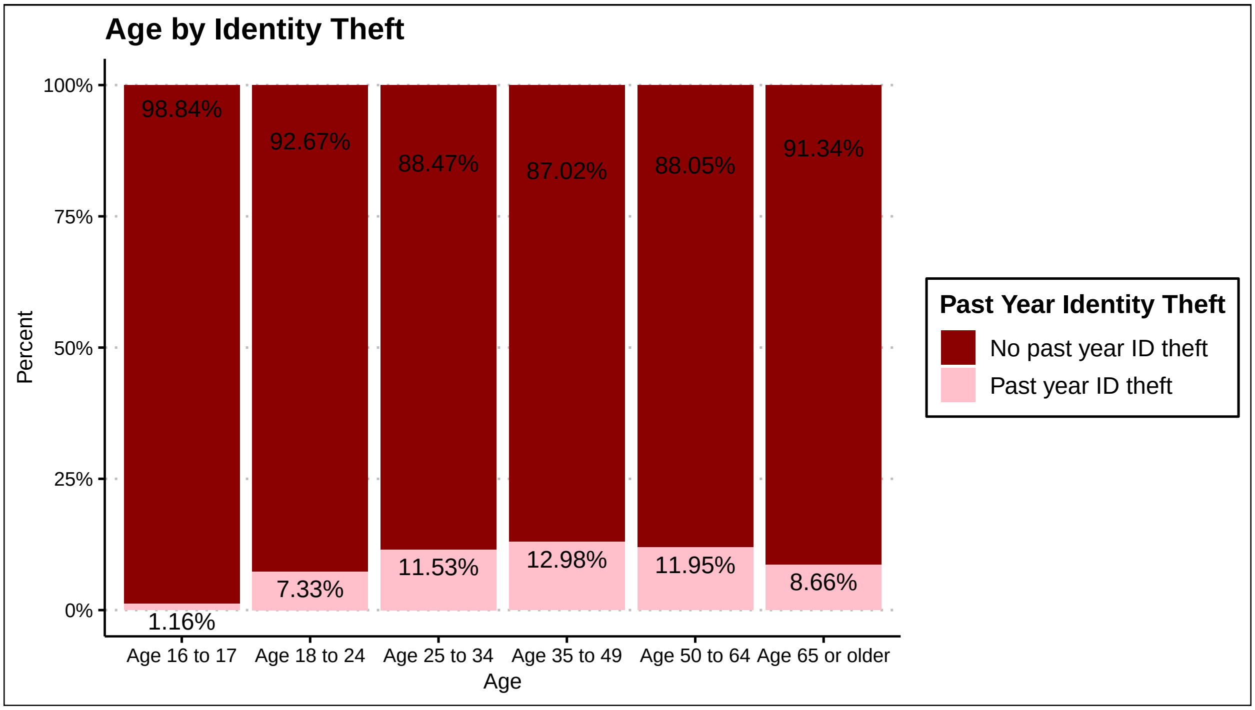
```
##                      Count Percentage
## Total                96130        100
## NonHispanic White    68265         71
## Not NonHispanic White 27865        29
```



Age

Twenty eight percent of the sample was age 50 to 64 while nearly one in four (24%) were age 35 to 49. 23% of the sample was age 65 or older. The remainder of the sample was under the age of 35. Thirteen percent of persons age 35 to 49 reported experiencing past year identity theft, compared to 7% of persons age 18 to 34 and 9% of persons age 65 or older.

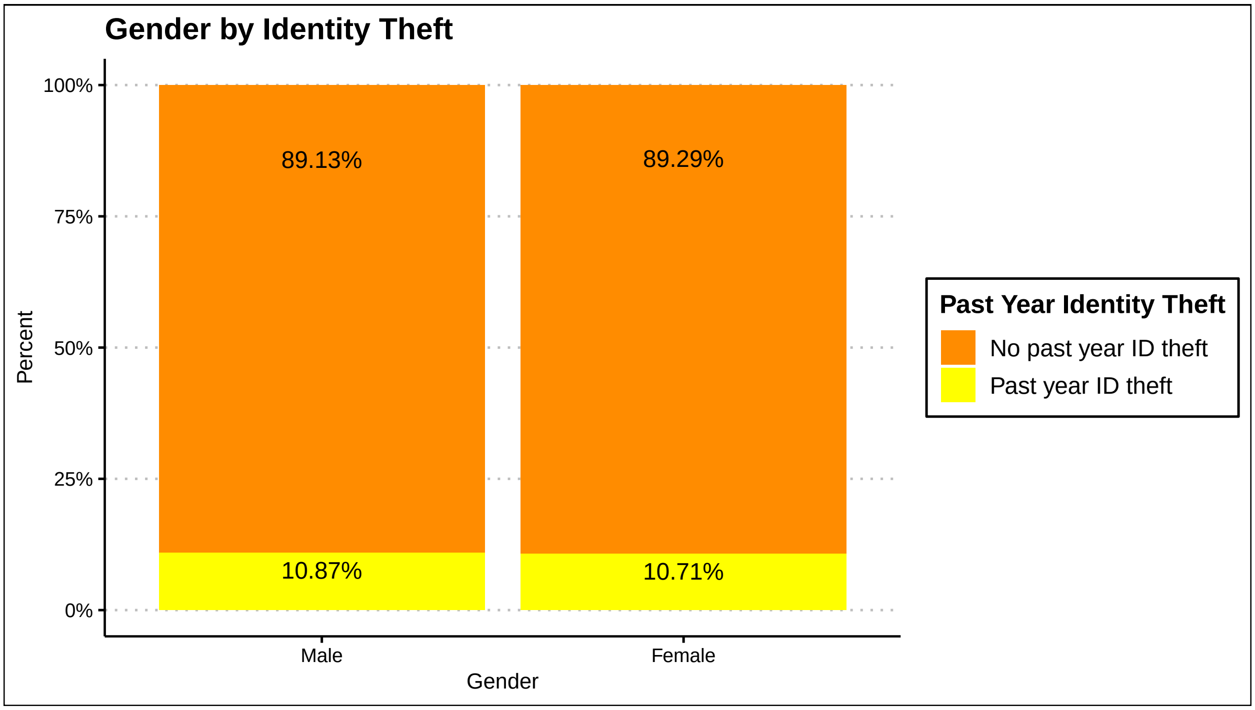
```
##                 Count Percentage
## Total           96130        100
## Age 16 to 17     1986          2
## Age 18 to 24     7826          8
## Age 25 to 34    14740         15
## Age 35 to 49    23147         24
## Age 50 to 64    26621         28
## Age 65 or older 21810         23
```



Gender

More than half of the sample (53%) was female while the remainder (47%) was male. Past year identity theft was experienced by 11% of males and a similar percentage of females.

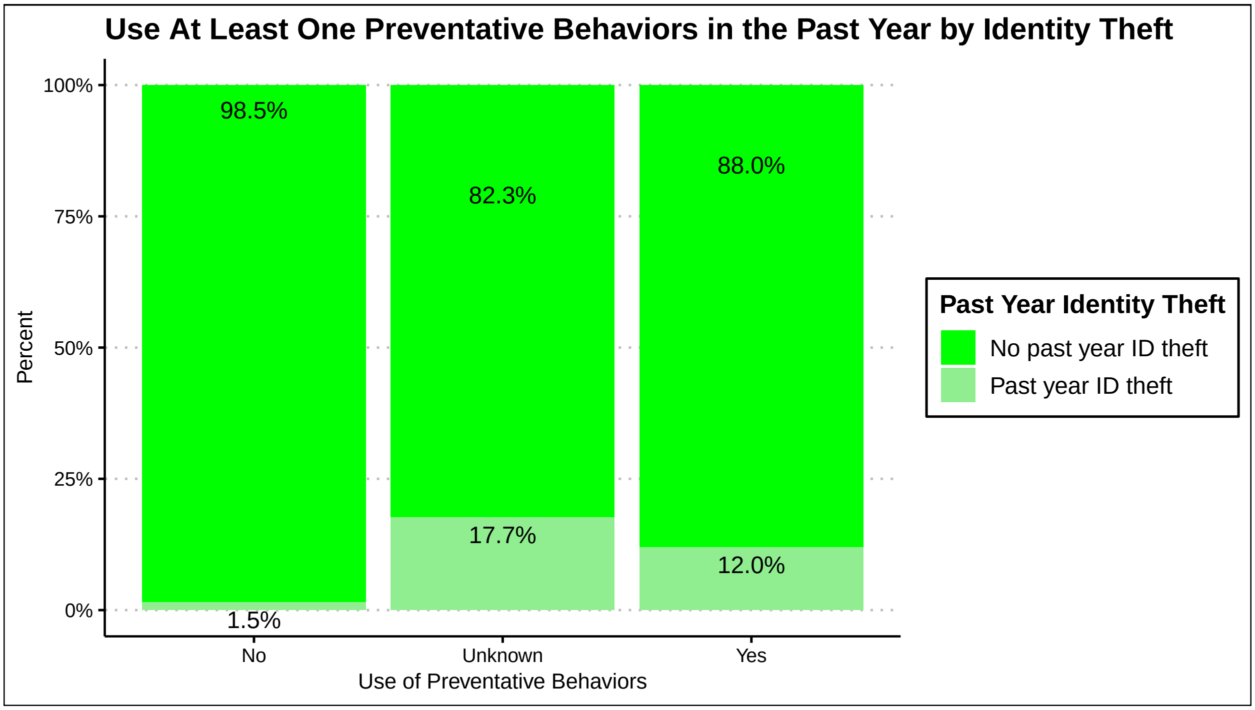
```
##          Count Percentage
## Total    96130        100
## Male     44908         47
## Female   51222         53
## Unknown      0          0
```



Preventative behaviors

Nearly nine out of ten persons in the sample (88%) used at least one of the preventative behaviors measured (checked credit report, changed password on financial accounts, had credit monitoring services or identity theft insurance, shredded or destroyed documents containing personal information, checked bank or credit card statements for unfamiliar charges, purchased identity theft protection) in the past 12 months. Eighteen percent of respondents who did not know if they had used a preventative behavior in the past 12 months reported past year identity theft compared to 12% of those who had used at least 1 preventative behavior in the past 12 months. Surprisingly, only 2% of those who did not participate in any prventative behaviors in the past 12 months reported past year identity theft.

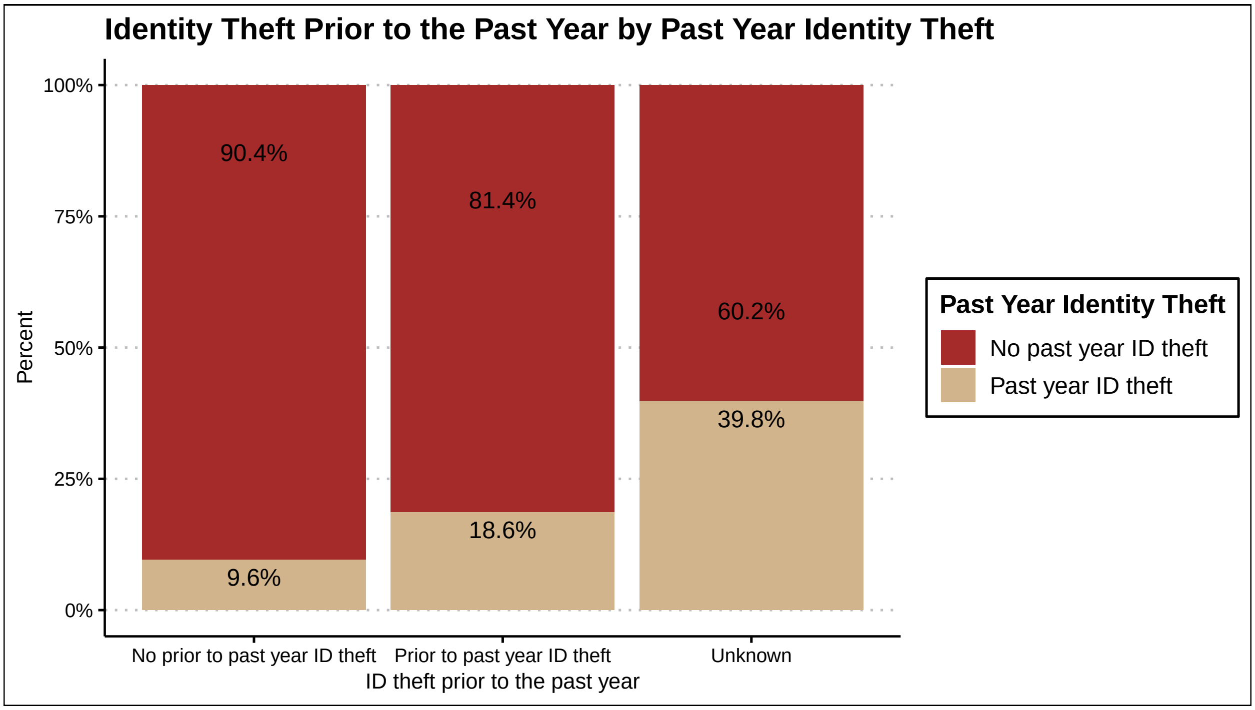
```
##          Count Percentage
## Total    96130        100
## No       11066         12
## Unknown    368          0
## Yes      84696         88
```



Identity theft prior to the past year

Thirteen percent of the sample experienced identity theft (misuse of an existing account, misuse of personal information to create new account or misuse of personal information for other fraudulent purposes) prior to 12 months prior to their ITS interview. The majority of the sample did not experience it. About 40% of respondents who did not know if they were a victim of identity theft prior to the past year experienced identity theft in the past 12 months. This is compared to 10% of those who had no identity theft prior to the past year and 19% of those who had identity theft prior to the past year.

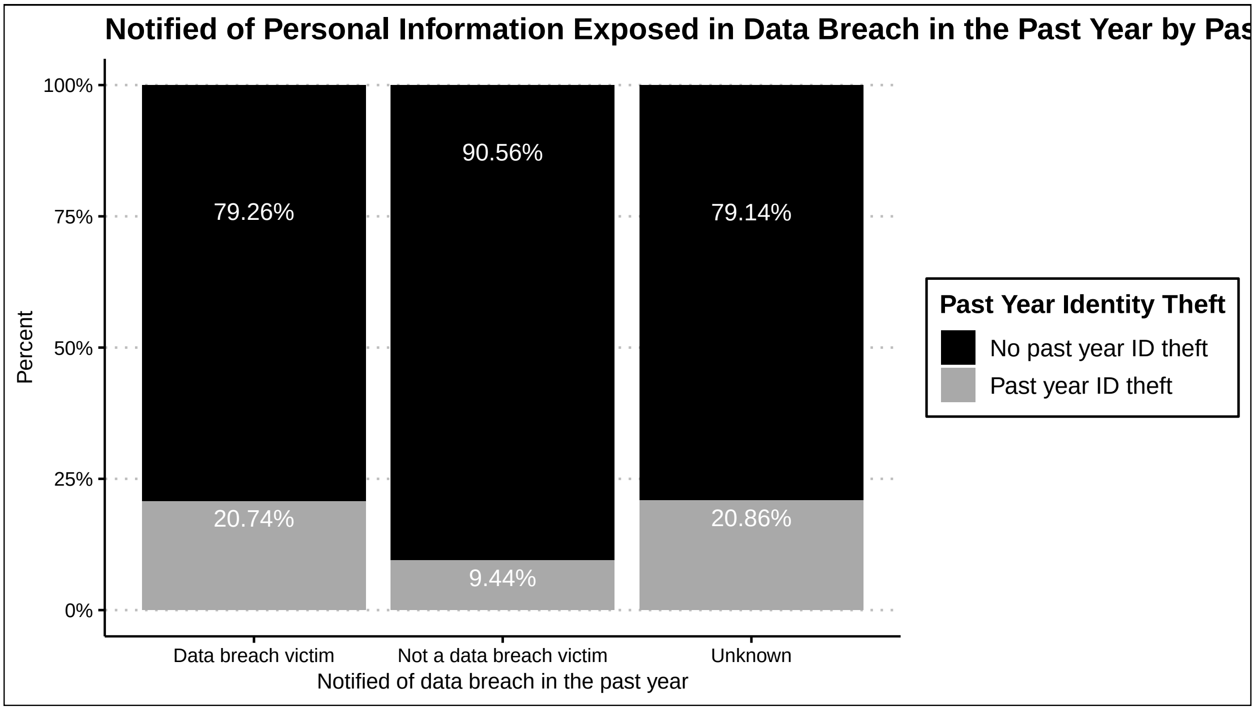
```
##                              Count Percentage
## Total                        96130        100
## No prior to past year ID theft 83692       87
## Prior to past year ID theft  12267         13
## Unknown                        171          0
```



Notified of exposure due to data breach

Twelve percent of the sample reported that they were notified that their personal information was exposed during a data breach. The majority of the sample (88%) reported that they were not notified that their personal information was exposed during a data breach.Of those who were notified that their information was exposed during a data breach, one in five (21%) reported being victims of identity theft in the past year. This is compared to 9% of those who were not notified that their personal information was exposed in a data breach being victims of past year identity theft.

```
##                          Count Percentage
## Total                    96130        100
## Data breach victim       11037         11
## Not a data breach victim 84652         88
## Unknown                    441          0
```



Data analysis

More data wrangling

Make copies of each variable used in analysis. Unknown level on each individual variable was changed to NA. Individual variables were combined into a single dataset and deleted individual variables. The individual variables were combined into a dataset (its1). The number of complete (cases with no NA values on any variable) and incomplete cases (cases with at least one variable with a value of NA) was summed from the dataset. There were 614 cases with a value of NA on at least one variable with the remaining cases being complete cases

```
## 'data.frame':    96130 obs. of  8 variables:
##  $ idtheft           : Factor w/ 2 levels "No past year ID theft ",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ incomer           : Factor w/ 2 levels "Less than $75,000",..: 2 2 2 2 1 1 1 1 1 1 ...
##  $ ager              : Factor w/ 6 levels "Age 16 to 17",..: 4 5 2 6 5 3 3 5 5 6 ...
##  $ ethnicr           : Factor w/ 2 levels "NonHispanic White",..: 1 2 2 1 1 1 1 1 1 1 ...
##  $ prevent_total     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ OUTSIDE_PAST_YEARR: Factor w/ 2 levels "No prior to past year ID theft",..: 1 1 2 1 1 2 1 1 1 1 .
##  $ notify_breachr    : Factor w/ 2 levels "Data breach victim",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ sexr              : Factor w/ 2 levels "Male","Female": 2 1 1 1 1 2 1 1 2 2 ...
```

```
##                    Number Percent
## Total               96130     100
## Cases with NAs        614       1
## Cases without NAs   95516      99
```

Chi-Square Analysis

Multiple chi-square analyses were run on the dataset with only completed cases which meant that 614 cases were dropped from data analysisi, leaving 95,516 cases to analyze. The chi-square analyses show that between past year identity theft was dependent on most of the predictors ($p < 0.05$) with the exception of sex ($p > 0.05$).

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  its1$idtheft and its1$incomer
## X-squared = 962.27, df = 1, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  its1$idtheft and its1$ager
## X-squared = 544.24, df = 5, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  its1$idtheft and its1$sexr
## X-squared = 0.39041, df = 1, p-value = 0.5321
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  its1$idtheft and its1$ethnicr
## X-squared = 359.61, df = 1, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  its1$idtheft and its1$prevent_total
## X-squared = 1117.8, df = 1, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  its1$idtheft and its1$OUTSIDE_PAST_YEARR
## X-squared = 899.98, df = 1, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  its1$idtheft and its1$notify_breachr
## X-squared = 1291.3, df = 1, p-value < 2.2e-16
```

Machine Learning

Setup

The dataset was split into training and testing datasets to in an attempt to train a logistics regression model predicting past year identity theft. Sixty percent of data (57,310 cases) was assigned to the training data while the remaining 40% (38,206 cases) went in the test dataset. The training dataset was examined for variables that had near zero or zero variability. There were no variables with zero or near zero variability.

```
##                    freqRatio percentUnique zeroVar   nzv
## idtheft             8.311129   0.003489792   FALSE FALSE
## incomer             1.855790   0.003489792   FALSE FALSE
## ager                1.149528   0.010469377   FALSE FALSE
## ethnicr             2.420267   0.003489792   FALSE FALSE
## prevent_total       7.633625   0.003489792   FALSE FALSE
## OUTSIDE_PAST_YEARR  6.748783   0.003489792   FALSE FALSE
## notify_breachr      7.695190   0.003489792   FALSE FALSE
## sexr                1.139551   0.003489792   FALSE FALSE
```

Training Model

A logistic regression model predicting past year identity theft was trained to the data. All of the predictors with the exception of sex was included in the model. Sex was excluded due to the chi-square analysis showing that sex was independent of past year identity theft. The results of the logistic regression model showed that all predictors were statistically significant in predicting past year identity theft (p<.05) with the exception of teh comparison of those age 25 to 34 to the reference group, persons age 35 to 49 (p>.05). Based on the odds ratios, the model showed that the odds of a person in households with annual incomes of $75,000 or more was 0.1 higher than that of a person in a household with annual incomes of less than $75,000. Each of the age groups tested had a lower odds of being a victim of past year identity theft than those in the reference group, age 35 to 49. The odds of a Non-Hispanic White person being a victim of past year identity theft was 1.3 times that of other persons. Being a victim of identity theft prior to the past year increased the odds of being a victim of identity theft in the past 12 months by a factor of about 1.6 over other persons. Being notified of personal information being exposed in a data breach increased the odds of being a victim of pst year identity theft by a factor of 1.7. One strange result from the model was that not using any of the preventative measures in the past 12 months reduced the odds of being a victim of past year identity theft by a factor of 0.2.

```
modFit <- glm(idtheft ~incomer+ager+ethnicr+prevent_total+
                OUTSIDE_PAST_YEARR+notify_breachr,data=training,family="binomial")
summary(modFit)
```

```
##
## Call:
## glm(formula = idtheft ~ incomer + ager + ethnicr + prevent_total +
##     OUTSIDE_PAST_YEARR + notify_breachr, family = "binomial",
##     data = training)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.8941  -0.5162  -0.4458  -0.3714   3.4247
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                  -2.39485    0.03863 -61.997
## incomer$75,000 or more                        0.42063    0.02839  14.815
## agerAge 16 to 17                             -1.64989    0.27260  -6.053
## agerAge 18 to 24                             -0.24466    0.06222  -3.932
## agerAge 25 to 34                             -0.07180    0.04281  -1.677
## agerAge 50 to 64                             -0.13471    0.03605  -3.737
## agerAge 65 or older                          -0.37815    0.04159  -9.092
## ethnicrNonHispanic White                      0.27068    0.03410   7.938
## prevent_totalNo                              -1.81676    0.10488 -17.322
## OUTSIDE_PAST_YEARRPrior to past year ID theft  0.44292    0.03500  12.656
## notify_breachrData breach victim              0.55000    0.03563  15.435
##                                              Pr(>|z|)
## (Intercept)                                   < 2e-16 ***
## incomer$75,000 or more                        < 2e-16 ***
## agerAge 16 to 17                             1.43e-09 ***
## agerAge 18 to 24                             8.42e-05 ***
## agerAge 25 to 34                             0.093499 .
## agerAge 50 to 64                             0.000186 ***
## agerAge 65 or older                           < 2e-16 ***
## ethnicrNonHispanic White                     2.06e-15 ***
## prevent_totalNo                               < 2e-16 ***
## OUTSIDE_PAST_YEARRPrior to past year ID theft  < 2e-16 ***
## notify_breachrData breach victim              < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39090  on 57309  degrees of freedom
## Residual deviance: 36956  on 57299  degrees of freedom
## AIC: 36978
##
## Number of Fisher Scoring iterations: 7
```

```
#Odds ratios for training model
exp(modFit$coeff)
```

```
##                                    (Intercept)
##                                     0.09118608
##                         incomer$75,000 or more
##                                     1.52291403
##                               agerAge 16 to 17
##                                     0.19207016
##                               agerAge 18 to 24
##                                     0.78296980
##                               agerAge 25 to 34
##                                     0.93071443
##                               agerAge 50 to 64
##                                     0.87397017
```

```
##                           agerAge 65 or older
##                                    0.68512600
##                     ethnicrNonHispanic White
##                                    1.31086181
##                               prevent_totalNo
##                                    0.16255160
## OUTSIDE_PAST_YEARRPrior to past year ID theft
##                                    1.55724819
##             notify_breachrData breach victim
##                                    1.73325236
```

Training Model Diagnostics

Variance inflation factors were generated to determine if there was multicollinearity among the predictors in the training model. The variance inflaction factors for all predictors in the model were about 1, showing that there was little to no multicollinearity among the predictors. An Anova model was fit to the training model to determine which factors were predictive in the model. The result of this analysis showed that all predictors were predictive (p<.05).

```
#variance inflation factors for training model
vif(modFit)
```

```
##                        GVIF Df GVIF^(1/(2*Df))
## incomer            1.063786  1        1.031400
## ager               1.075482  5        1.007303
## ethnicr            1.050755  1        1.025063
## prevent_total      1.019679  1        1.009791
## OUTSIDE_PAST_YEARR 1.054309  1        1.026795
## notify_breachr     1.067470  1        1.033184
```

```
#anova- used for putting factors in and out of training model
anova(modFit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: idtheft
##
## Terms added sequentially (first to last)
##
##
##                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                               57309      39090
## incomer             1   574.92     57308      38515 < 2.2e-16 ***
## ager                5   314.48     57303      38201 < 2.2e-16 ***
## ethnicr             1   169.35     57302      38031 < 2.2e-16 ***
## prevent_total       1   630.28     57301      37401 < 2.2e-16 ***
## OUTSIDE_PAST_YEARR  1   220.93     57300      37180 < 2.2e-16 ***
## notify_breachr      1   223.68     57299      36956 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running Adjusted Model

Another model was ran that was the original training model without the predictor for the use of preventative behaviors in the past 12 months. This was done due to the unusual result in the original training model. Similar results were obtained. The majority of the predictors were found to be statistically significant (p<.05). The only exception was that the comparison of persons age 35 to 45 to those age 25 to 34 was not statistically significant (p>0.5). According to the odds ratios for the adjusted model, the odds of persons in households with annual incomes or $75,000 or more being a victim of past year identity theft increased by 0.1 compared to other persons. Each of the age groups tested had a lower odds of being a victim of past year identity theft than those in the reference group, age 35 to 49. The odds of a nonHispanic White person being a victim of identity theft was 1.4 times the odds of a person of another racial/ethnic group being a victim.Being a victim of identity theft prior to the past year increased the odds of being a victim during the past 12 months by 1.6 compared to other persons. Being notified that personal information was exposed during a data brief increased the odds of being a victim of past year identity theft by 1.9 compared to those who were not notified of being a victim of a data breach.

```r
modFit1 <- glm(idtheft ~ incomer+ager+ethnicr++OUTSIDE_PAST_YEARR+notify_breachr
              ,data=training,family="binomial")
#get model estimates for adjusted model
summary(modFit1)
```

```
##
## Call:
## glm(formula = idtheft ~ incomer + ager + ethnicr + +OUTSIDE_PAST_YEARR +
##     notify_breachr, family = "binomial", data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9151  -0.5047  -0.4263  -0.3692   3.1246
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                              -2.58097    0.03803 -67.875
## incomer$75,000 or more                    0.47053    0.02834  16.600
## agerAge 16 to 17                         -2.29296    0.27001  -8.492
## agerAge 18 to 24                         -0.31912    0.06183  -5.161
## agerAge 25 to 34                         -0.07067    0.04268  -1.656
## agerAge 50 to 64                         -0.12862    0.03597  -3.576
## agerAge 65 or older                      -0.37975    0.04146  -9.159
## ethnicrNonHispanic White                  0.35697    0.03393  10.521
## OUTSIDE_PAST_YEARRPrior to past year ID theft  0.49379    0.03510  14.070
## notify_breachrData breach victim          0.60571    0.03575  16.942
##                                          Pr(>|z|)
## (Intercept)                               < 2e-16 ***
## incomer$75,000 or more                    < 2e-16 ***
## agerAge 16 to 17                          < 2e-16 ***
## agerAge 18 to 24                         2.45e-07 ***
## agerAge 25 to 34                         0.097773 .
## agerAge 50 to 64                         0.000349 ***
## agerAge 65 or older                       < 2e-16 ***
## ethnicrNonHispanic White                  < 2e-16 ***
```

```
## OUTSIDE_PAST_YEARRPrior to past year ID theft  < 2e-16 ***
## notify_breachrData breach victim              < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39090  on 57309  degrees of freedom
## Residual deviance: 37489  on 57300  degrees of freedom
## AIC: 37509
##
## Number of Fisher Scoring iterations: 6
```

```
#Odds ratios for adjusted model
exp(modFit1$coeff)
```

```
##                                  (Intercept)
##                                   0.07570053
##                         incomer$75,000 or more
##                                   1.60084527
##                            agerAge 16 to 17
##                                   0.10096706
##                            agerAge 18 to 24
##                                   0.72679050
##                            agerAge 25 to 34
##                                   0.93176583
##                            agerAge 50 to 64
##                                   0.87931163
##                         agerAge 65 or older
##                                   0.68402953
##                       ethnicrNonHispanic White
##                                   1.42898826
## OUTSIDE_PAST_YEARRPrior to past year ID theft
##                                   1.63850882
##                 notify_breachrData breach victim
##                                   1.83255814
```

Adjusted model diagnostics

Variance inflation factors were generated on the adjusted model and showed little to no multicollinearity among the predictors. The Anova chi-square results showed that each of the predictors were predictive in the model.

```
#variance inflation factors for adjusted model
vif(modFit1)
```

```
##                        GVIF Df GVIF^(1/(2*Df))
## incomer            1.065277  1        1.032122
## ager               1.067071  5        1.006513
## ethnicr            1.051348  1        1.025353
## OUTSIDE_PAST_YEARR 1.058836  1        1.028998
## notify_breachr     1.072691  1        1.035708
```

```r
#anova- used for putting factors in and out of adjusted model
anova(modFit1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: idtheft
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                57309      39090
## incomer              1   574.92     57308      38515 < 2.2e-16 ***
## ager                 5   314.48     57303      38201 < 2.2e-16 ***
## ethnicr              1   169.35     57302      38031 < 2.2e-16 ***
## OUTSIDE_PAST_YEARR  1   274.79     57301      37757 < 2.2e-16 ***
## notify_breachr       1   267.69     57300      37489 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing both models

```r
anova(modFit,modFit1,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: idtheft ~ incomer + ager + ethnicr + prevent_total + OUTSIDE_PAST_YEARR +
##     notify_breachr
## Model 2: idtheft ~ incomer + ager + ethnicr + +OUTSIDE_PAST_YEARR + notify_breachr
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     57299      36956
## 2     57300      37489 -1  -532.42 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Accuracy of both models

```r
pred<-predict(modFit,newdata=training,type="response")
pred<-ifelse(pred>0.5,1,0)
pred1<-as.vector(predict(modFit1,newdata=training,type="response"))
pred1<-ifelse(pred1>0.5,1,0)

pr <- prediction(pred,training$idtheft)
perf <- performance(pr,measure = "tpr",x.measure = "fpr")
```

```r
#plot ROC
#plot(perf, title("ROC for training model"))
#print AUC score for training
auc(training$idtheft,pred)
```

```
## Setting levels: control = No past year ID theft , case = Past year ID theft

## Setting direction: controls < cases

## Area under the curve: 0.5
```

```r
pr1 <- prediction(pred1,training$idtheft)
perf1 <- performance(pr1,measure = "tpr",x.measure = "fpr")
#plot ROC
#plot(perf1, title("ROC for adjusted model"))
#print AUC score for adjusted model
auc(training$idtheft,pred1)
```

```
## Setting levels: control = No past year ID theft , case = Past year ID theft
## Setting direction: controls < cases

## Area under the curve: 0.5
```