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

Erika Harrell

August 2021

Disclaimer: The points and views expressed in this document do not necessarily reflect the views of the Bureau of Justice Statistics nor that of the U.S. Department of Justice.

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). Chi-square results 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 (p> .05). Three logistic regression models were trained on the data, including one in which the outcome variable under went undersampling of cases where the respondent did not report past year identity theft. However, the best model was one without gender or preventative behaviors as predictors, with no undersampling of the outcome. This model had higher precision and accuracy than the other 2 models.

```
#adding libraries
library(reshape2)
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(caTools)
library(pROC)
```

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

```
##
## Attaching package: 'pROC'
```

```
## 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 and cleaning

Reducing the dataset- There were 125,165 total persons in the 2016 ITS dataset. Just using the completed telephone and personal interviews leaves 96,130 interviews or observations in the dataset.

```
##                          Count Percentage
## (1) Personal interview   57119         46
## (2) Telephone interview  39011         31
## (5) ITS Noninterview     29035         23
## Total                   125165        100
```

```
##                         Count Percentage
## (1) Personal interview  57119         59
## (2) Telephone interview 39011         41
## (5) ITS Noninterview        0          0
## Total                   96130        100
```

From the larger ITS dataset, several variables were used: past year identity theft (misuse or attempted misuse of an existing account, misuse or attempted misuse of personal inforamtion to opena new account of for some other fraudulent purpose) age, race/Hispanic origin, gender, annual household income, whether a respondet participated in behaviors in the past year to prevent identity theft, whether a respondent experienced identity theft prior to the past year, whther a respondent was notified that their personal information was exposed due to a data breach.

Individual variables creaetd 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
##
##
```

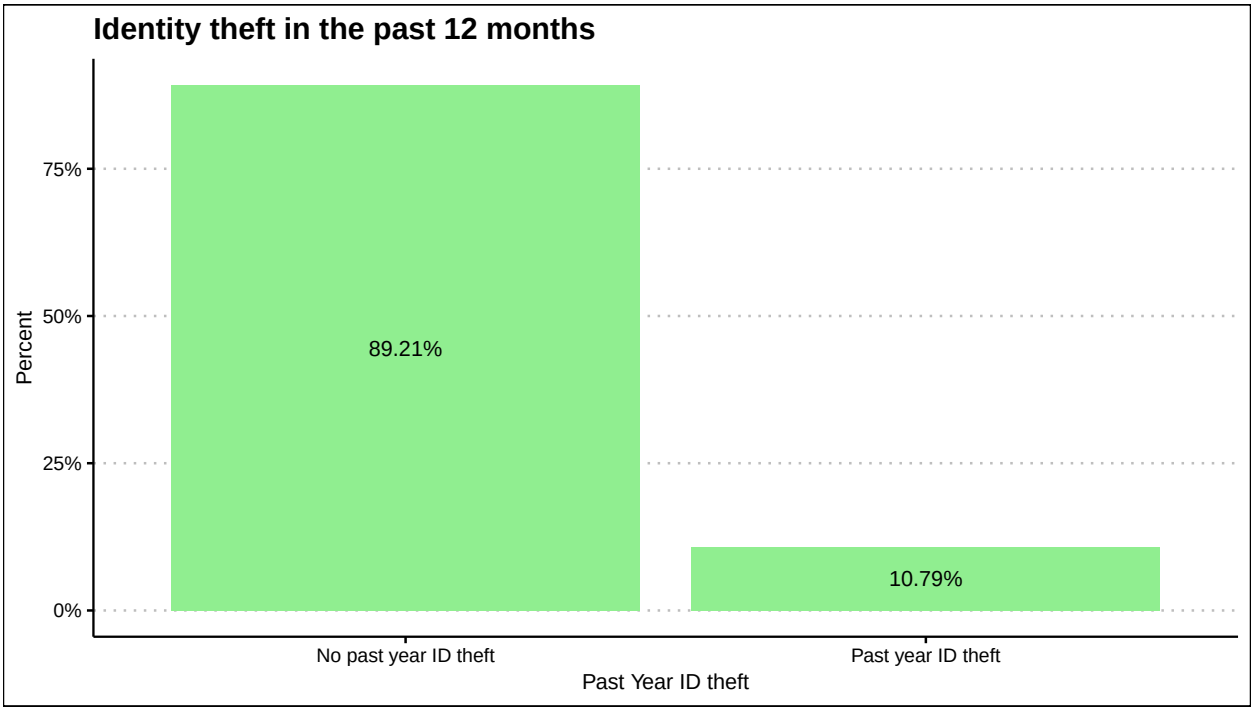
Variables were recoded, collapsing categories and computing necessary categories and variables for analysis.

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 in the past year.

```
##                        Count Percentage
## No past year ID theft 85762         89
## Past year ID theft    10368         11
## Total                 96130        100
```

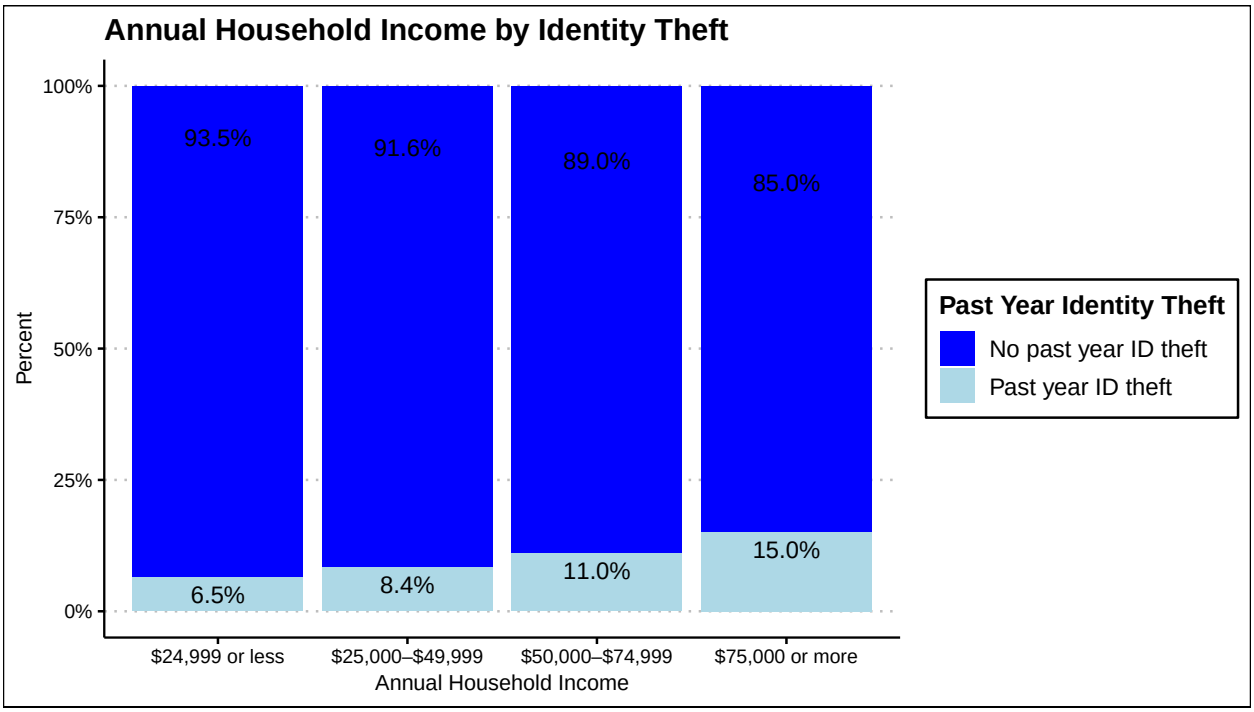


Annual household income

One in five (21%) respondents were in households with annual incomes of $24,999 or less. Another fourth (26%) of the sample were in households with annual incomes of $25,000 to $49,999. Eighteen percent (18%) of the sample were in households with annual incomes of $50,000 to $74,999. The reminder of the sample (35%) were in households with annual incomes of at least $75,000. Identity theft affected a higher percentage of those in households with higher annual incomes. Fifteen percent (15%) of persons in households with annual incomes of at least $75,000 reported experiencing identity theft in the past 12 months compared to 6.5% of those in household with annual incomes of $24,999 or less.

```
##                  Count Percentage
```

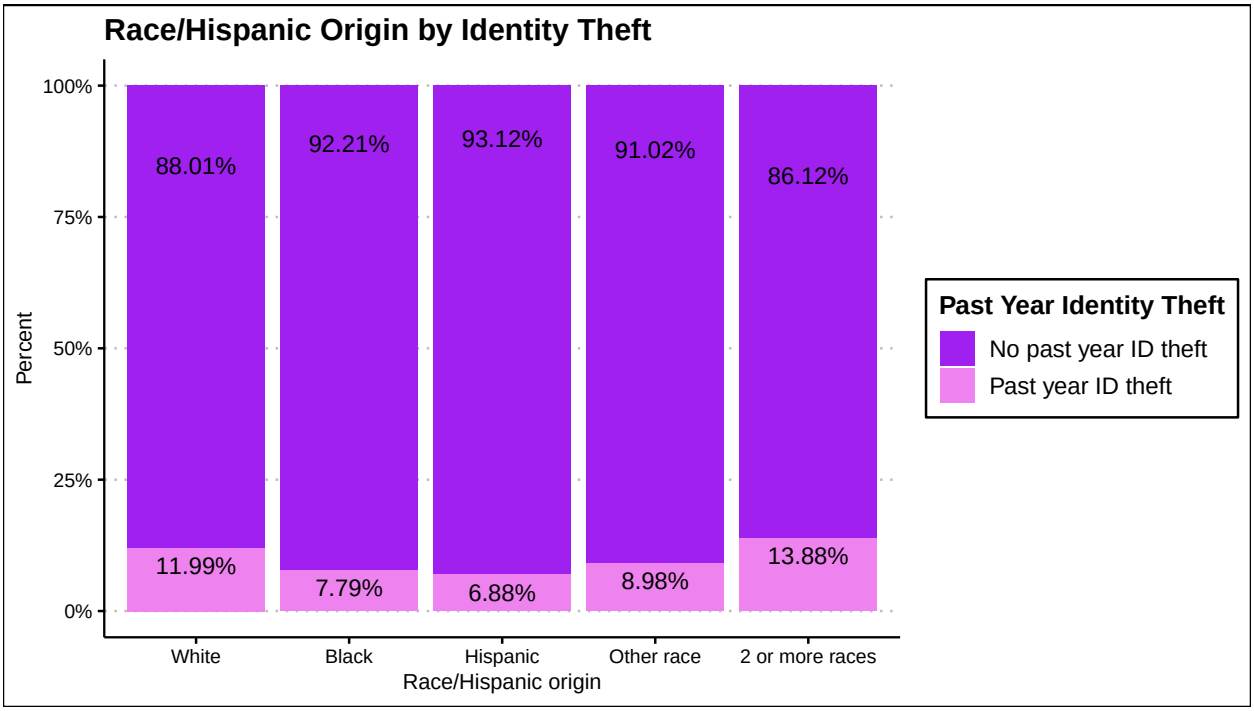
```
## $24,999 or less 19871         21
## $25,000-$49,999 25255         26
## $50,000-$74,999 17342         18
## $75,000 or more 33662         35
## Total           96130        100
```



Race/Hispanic origin

Seventy-one percent (71%) of the sample were White while one in ten (10%) respondents were Black. Hispanics accounted for 13% of respondents. Persons who were of another race accounted for 5% of the sample. Persons of 2 or more races accounted for 1% of the sample. Identity theft appeared to account for a larger percentage of Whites (12%) and persons of 2 or more races (14%) than Blacks (8%), Hispanics (7%) and persons of other races (9%).

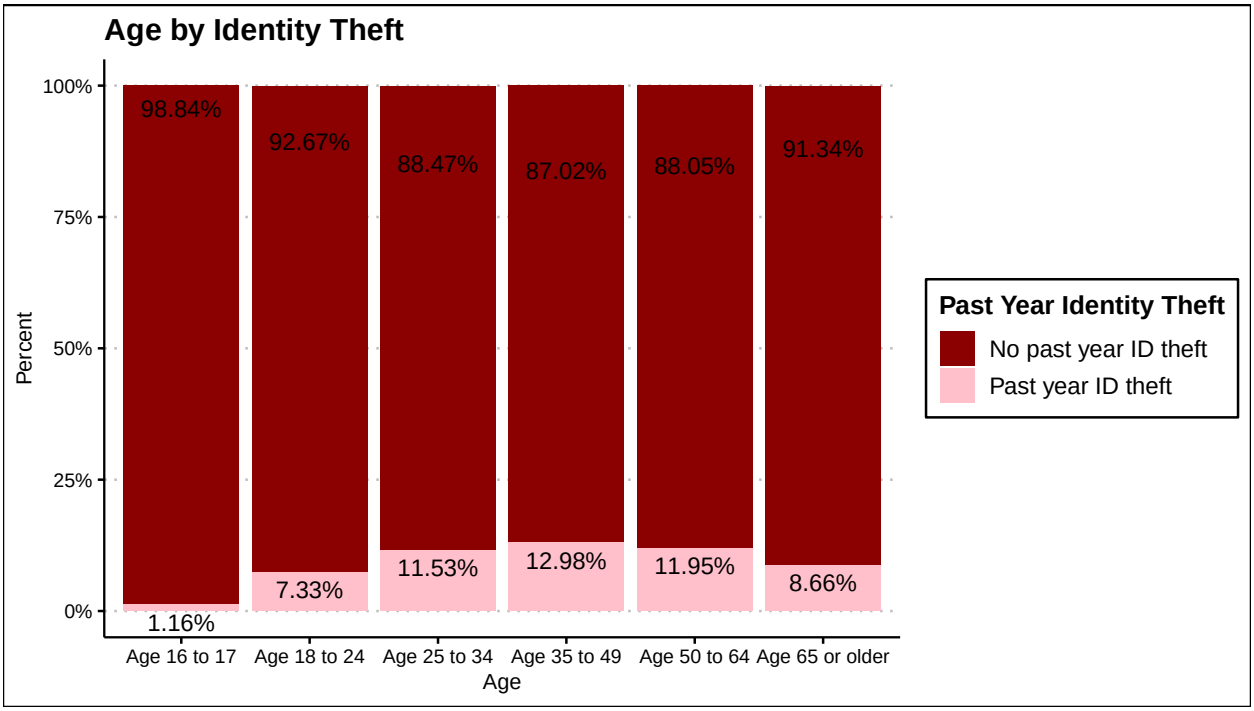
```
##                Count Percentage
## White          68265         71
## Black           9839         10
## Hispanic       12131         13
## Other race      4865          5
## 2 or more races 1030          1
## Total          96130        100
```



Age

Twenty-eight(28%) of the sample was age 50 to 64 while nearly one in four (24%) were age 35 to 49. Twenty-three percent (23%) of the sample was age 65 or older. The remainder of the sample was under the age of 35. Thirteen percent (13%) of persons age 35 to 49 reported past year identity theft, compared to 7% of persons age 18 to 34 and 9% of persons age 65 or older.

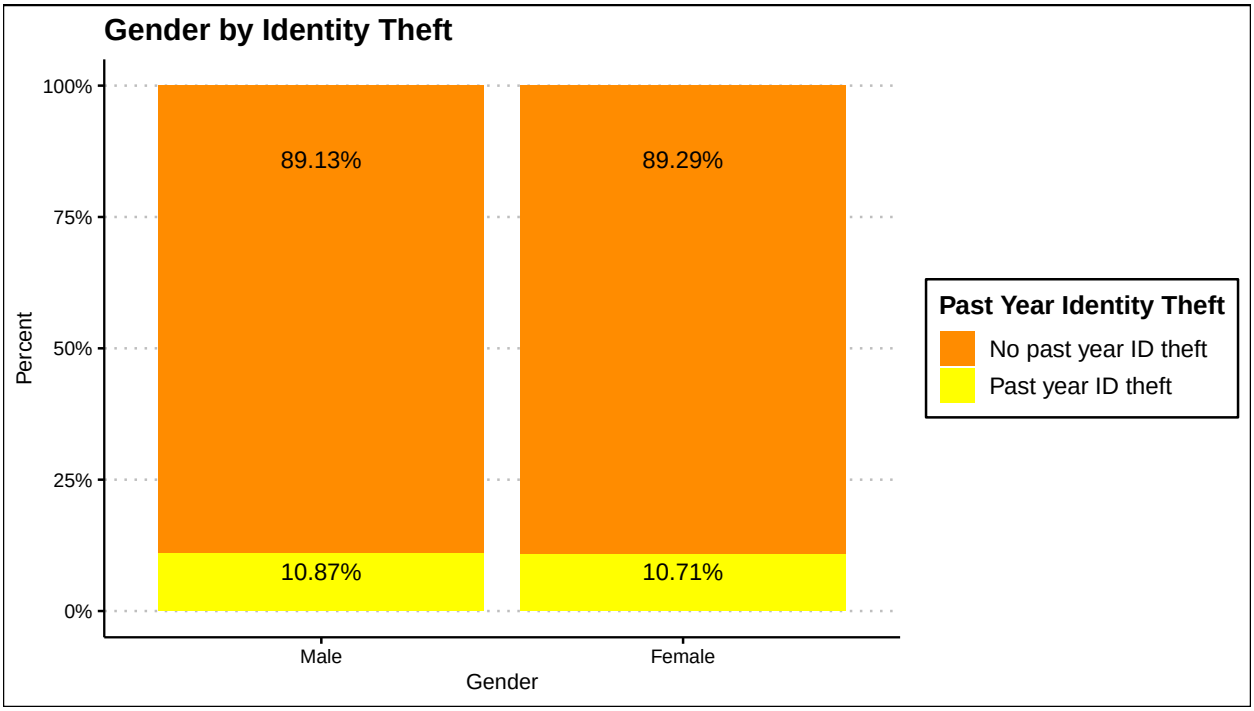
```
##                  Count Percentage
## Age 16 to 17      1986          2
## Age 18 to 24      7826          8
## Age 25 to 34     14740         15
## Age 35 to 49     23147         24
## Age 50 to 64     26621         28
## Age 65 or older  21810         23
## Total            96130        100
```



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.

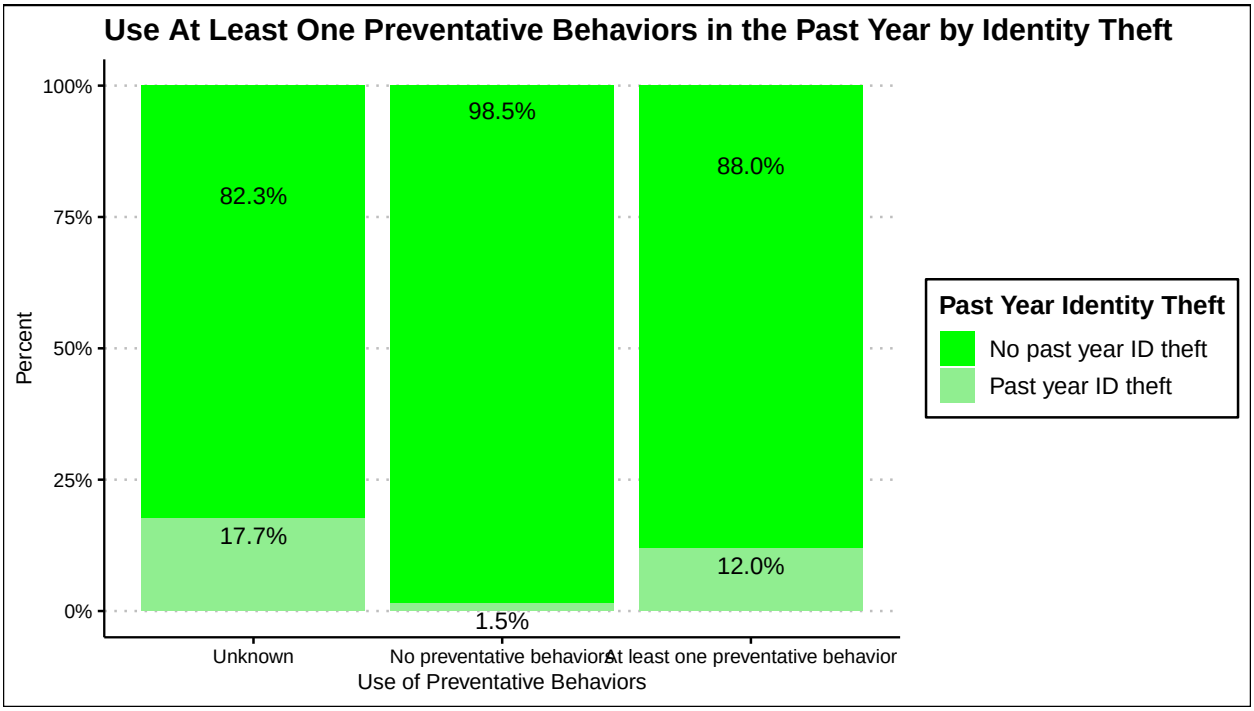
```
##        Count Percentage
## Male   44908         47
## Female 51222         53
## Total  96130        100
```



Preventative behaviors

Nearly nine out of ten respondents (88%) used at least one of the preventative behaviors measured (checked bank or credit card statements, shredded or destroyed documents with financial information, checked credit report, changed passwords on financial accounts, used identity-theft security program on computer, purchased identity theft insurance or credit monitoring service, purchased identity-theft protection) in the past 12 months to prevent being a victim of identity theft. Eighteen percent of persons 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 one preventative behavior in the past 12 months. Unusually, 2% of those who reported no preventative behaviors reported identity theft in the past year.

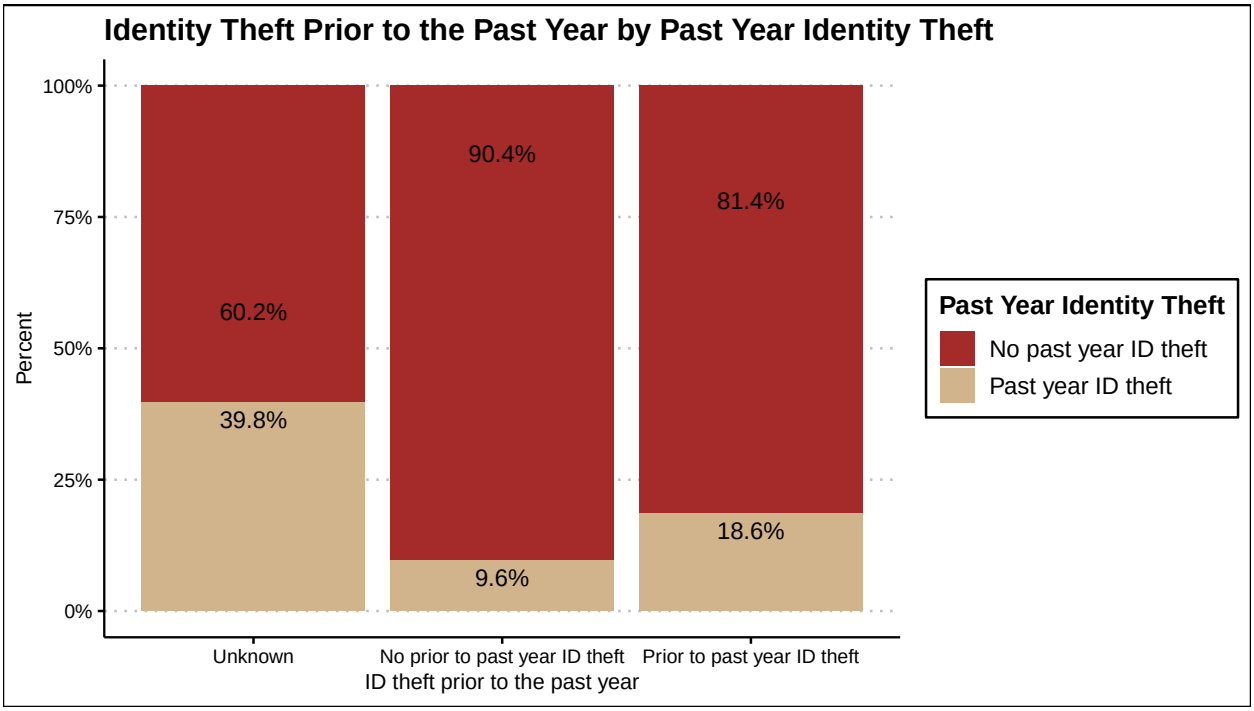
```
##                                      Count Percentage
## Unknown                                368          0
## No preventative behaviors            11066         12
## At least one preventative behavior   84696         88
## Total                                96130        100
```



Identity theft prior to the past year

Thirteen percent (13%) 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 the 12 months prior to their ITS interview. The majority of the sample (87%) did not experience it. About one in five (19%) respondents who experienced identity theft prior to the past year also experienced identity theft in the past 12 months. About 10% of those who did not have identity theft prior to the past 12 months experienced identity theft in the past 12 months.

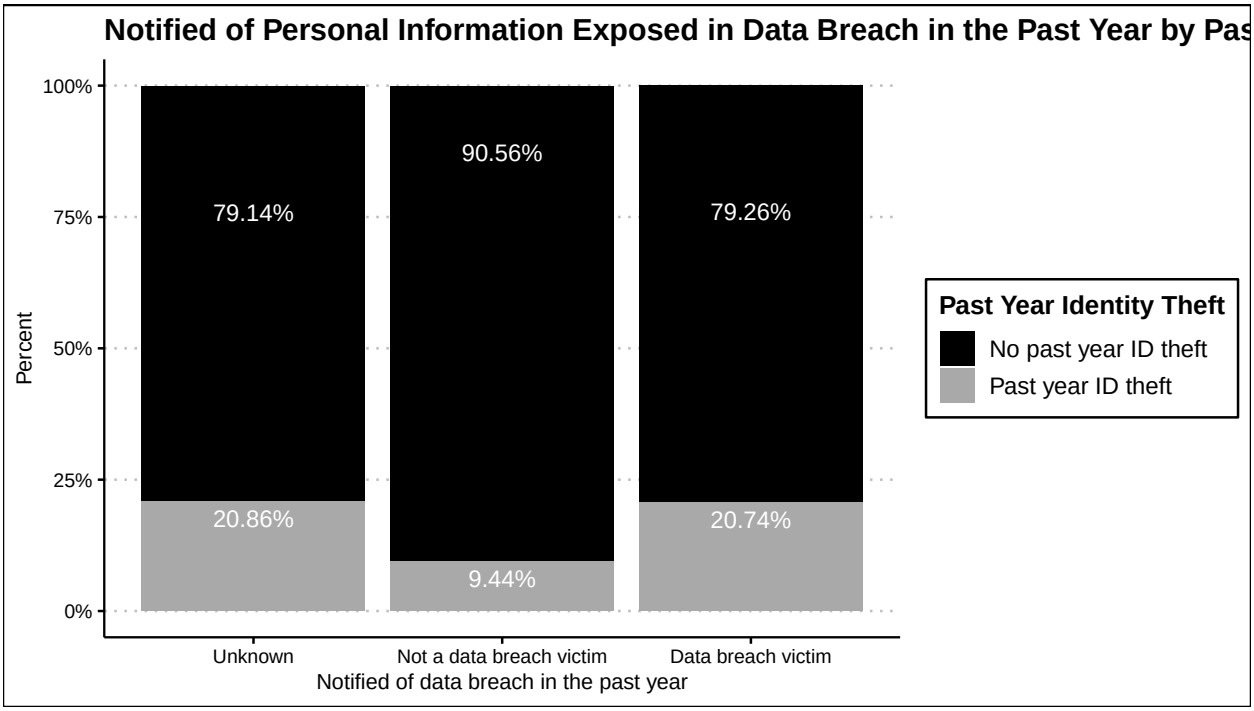
```
##                                Count Percentage
## Unknown                          171          0
## No prior to past year ID theft 83692         87
## Prior to past year ID theft    12267         13
## Total                          96130        100
```



Notified of personal information exposure due to data breach

Eleven percent (11%) 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
## Unknown                      441          0
## Not a data breach victim   84652         88
## Data breach victim         11037         11
## Total                      96130        100
```



Data analysis

More data wrangling

Unknown level on each individual variable used in the analysis was changed to NA. Individual variables were combined into a single dataset. 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 95,516 cases being complete cases with no missing values.

```
##                  Number Percent
## Incomplete cases    614       1
## Complete cases    95516      99
## Total             96130     100
```

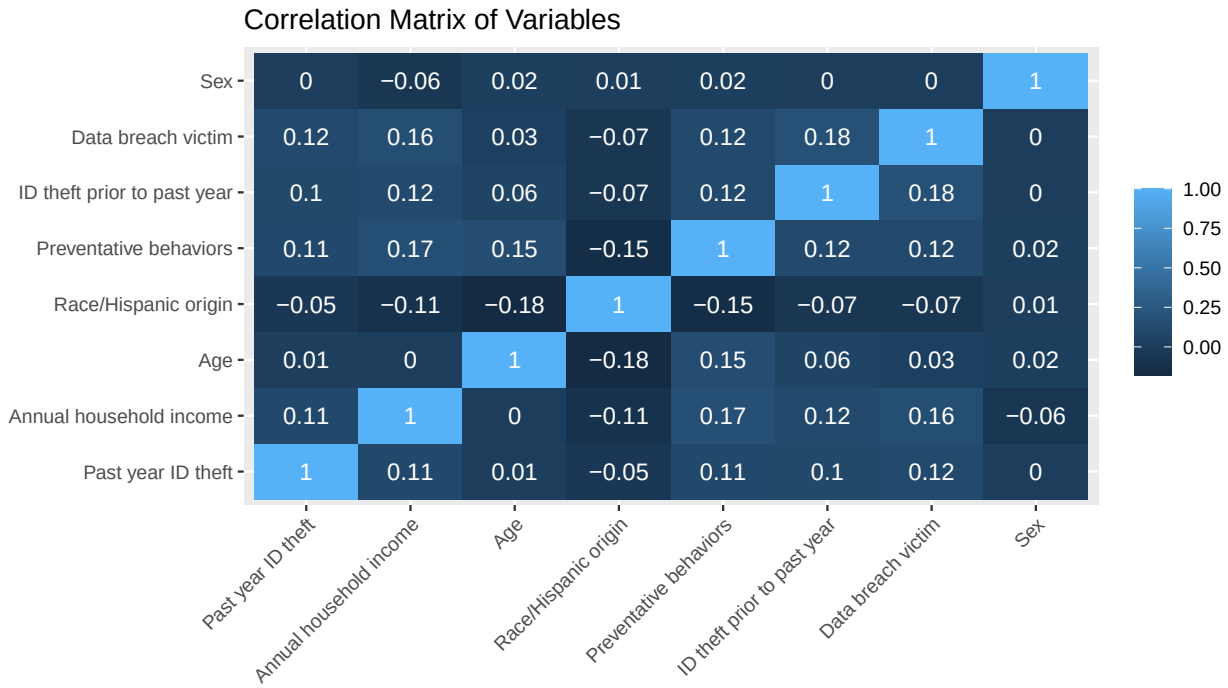
```
##                   idtheft                 incomer                   ager
##  No past year ID theft:85258   $24,999 or less:19746   Age 16 to 17   : 1975
##  Past year ID theft   :10258   $25,000-$49,999:25108   Age 18 to 24   : 7799
##                                $50,000-$74,999:17245   Age 25 to 34   :14649
##                                $75,000 or more:33417   Age 35 to 49   :22990
##                                                        Age 50 to 64   :26452
##                                                        Age 65 or older:21651
##             ethnicr                                  prevent_total
##  White          :67836   No preventative behaviors         :11051
##  Black          : 9777   At least one preventative behavior:84465
##  Hispanic       :12059
##  Other race     : 4822
##  2 or more races: 1022
##
##                       OUTSIDE_PAST_YEARR                 notify_breachr
```

```
##  No prior to past year ID theft:83294    Not a data breach victim:84506
##  Prior to past year ID theft   :12222    Data breach victim      :11010
##
##
##
##
##      sexr
##  Male  :44617
##  Female:50899
##
##
##
##
```

Correlation matrix

Pearson correlations of the variables were generated. The predictors appear to be relatively independent of each other with no moderate or strong correlations in the dataset. Gender appeared to have no correlation to the outcome variable, past year identity theft (r=0).

```
##                 Var1    Var2 value
## 1            idtheft idtheft  1.00
## 2            incomer idtheft  0.11
## 3               ager idtheft  0.01
## 4            ethnicr idtheft -0.05
## 5      prevent_total idtheft  0.11
## 6 OUTSIDE_PAST_YEARR idtheft  0.10
```



Chi-Square Analysis

Multiple chi-square analyses were run comparing each predictor to the outcome variable. 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
##
## data:  its_clean$idtheft and its_clean$incomer
## X-squared = 1157.7, df = 3, p-value < 2.2e-16
```

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

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

```
##
##  Pearson's Chi-squared test
##
## data:  its_clean$idtheft and its_clean$ethnicr
## X-squared = 417.79, df = 4, p-value < 2.2e-16
```

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

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

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

Machine learning

Models were trained onto the data in an attempt to predict past year identity theft. The data as split into a training and test dataset with 70% of cases in the train dataset and 30% of cases in the test dataset. A logistic regression model that contained all variables with the exception of gender (gender was deleted due to its lack of correlation with the outcome variable) was trained on the data. The resulting model showed that all of the predictors were statistically significant (p<.05) in predicting past year identity theft. Respondents who were in households with higher annual incomes, Whites and persons of 2 or more races, older persons,those who had experienced identity theft prior to the past 12 months, and those who were notified that there information was exposed in a data breach were more likely than others to report identity theft in the past year. One unusual finding in the model was that participation in at least one preventative measure was significant in being a victim of identity theft in the past year. This is addressed in the model below.The model was then assessed for accuracy, recall and precision. Its F1 score was also generated. The accuracy score was .58, meaning that the model only predicted 58% of cases correctly. Its precision was .56 meaning that when the model classified a respondent as a victim of identity theft in the past year, it was correct 56% of the time. Its recall was .94, meaning that it correctly idenitifed 94% of all victims of identity theft. Its F1 score was .70. Efforts were made to improve the model below.

```
##
## Call:
## glm(formula = idtheft ~ incomer + ager + ethnicr + OUTSIDE_PAST_YEARR +
##     prevent_total + notify_breachr, family = "binomial", data = trainSet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9687  -0.5269  -0.4329  -0.3421   3.2525
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                    -5.84538    0.27482 -21.270
## incomer$25,000-$49,999                          0.10310    0.04484   2.299
## incomer$50,000-$74,999                          0.29719    0.04640   6.405
## incomer$75,000 or more                          0.56108    0.04108  13.659
## agerAge 18 to 24                                1.44229    0.26731   5.396
## agerAge 25 to 34                                1.70565    0.26378   6.466
## agerAge 35 to 49                                1.73690    0.26304   6.603
## agerAge 50 to 64                                1.64286    0.26298   6.247
## agerAge 65 or older                             1.39959    0.26364   5.309
## ethnicrBlack                                   -0.24197    0.04923  -4.916
## ethnicrHispanic                                -0.32700    0.04689  -6.974
## ethnicrOther race                              -0.26480    0.06404  -4.135
## ethnicr2 or more races                          0.22492    0.11197   2.009
## OUTSIDE_PAST_YEARRPrior to past year ID theft   0.42626    0.03255  13.097
## prevent_totalAt least one preventative behavior 1.81809    0.09981  18.215
## notify_breachrData breach victim                0.56504    0.03276  17.249
##                                                Pr(>|z|)
## (Intercept)                                     < 2e-16 ***
## incomer$25,000-$49,999                           0.0215 *
## incomer$50,000-$74,999                         1.50e-10 ***
## incomer$75,000 or more                          < 2e-16 ***
## agerAge 18 to 24                               6.83e-08 ***
## agerAge 25 to 34                               1.01e-10 ***
```

```
## agerAge 35 to 49                                    4.03e-11 ***
## agerAge 50 to 64                                    4.18e-10 ***
## agerAge 65 or older                                 1.10e-07 ***
## ethnicrBlack                                        8.85e-07 ***
## ethnicrHispanic                                     3.08e-12 ***
## ethnicrOther race                                   3.55e-05 ***
## ethnicr2 or more races                                0.0446 *
## OUTSIDE_PAST_YEARRPrior to past year ID theft      < 2e-16 ***
## prevent_totalAt least one preventative behavior    < 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: 45606  on 66861  degrees of freedom
## Residual deviance: 42990  on 66846  degrees of freedom
## AIC: 43022
##
## Number of Fisher Scoring iterations: 7

## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 14428  1003
##          1 11149  2074
##
##                Accuracy : 0.5759
##                  95% CI : (0.5702, 0.5816)
##     No Information Rate : 0.8926
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0972
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5641
##             Specificity : 0.6740
##          Pos Pred Value : 0.9350
##          Neg Pred Value : 0.1568
##              Prevalence : 0.8926
##          Detection Rate : 0.5035
##    Detection Prevalence : 0.5385
##       Balanced Accuracy : 0.6191
##
##        'Positive' Class : 0
##

## [1] "The recall of the above model is 0.935000972069211."

## [1] "The precision of the above model is 0.564100559096063."

## [1] "The F1 score of the above model is 0.703667577058135."
```

Another logistic regression model was trained on the data. Checking the 2016 ITS questionnaire revealed that questions surrounding preventative behaviors were somewhat vague and did not specify a specific temporal order for the preventative behaviors and the identity theft victimization. It is possible that a respondent could have engaged in the preventative behaviors after becoming a victim as well as prior to becoming a victim. With this in mind, preventative behaviors was removed from the model and the model was trained on data that was again split (70% of cases in the train dataset/30% of cases in the test dataset) with a different seed set to protect against overfitting. In the model, again all predictors appeared to be statistically significant in predicting past year identity theft (p< .05) with similar results regarding comparisons as the previous model. The model was then used to predict the outcome in the test data and metrics used to assess the quality of the model were generated. The accuracy score (.64), F1 score (.78), and the precision (.65) improved while the recall (.93) declined. Another improvement to the data was made and assessed below.

```
##
## Call:
## glm(formula = idtheft ~ incomer + ager + ethnicr + OUTSIDE_PAST_YEARR +
##     notify_breachr, family = "binomial", data = trainSet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0152  -0.5239  -0.4212  -0.3423   3.1721
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                 -4.73407    0.24669 -19.191
## incomer$25,000-$49,999                       0.17787    0.04465   3.984
## incomer$50,000-$74,999                       0.43686    0.04587   9.524
## incomer$75,000 or more                       0.69730    0.04062  17.165
## agerAge 18 to 24                             1.91831    0.24964   7.684
## agerAge 25 to 34                             2.21548    0.24609   9.003
## agerAge 35 to 49                             2.24793    0.24530   9.164
## agerAge 50 to 64                             2.13961    0.24525   8.724
## agerAge 65 or older                          1.91964    0.24594   7.805
## ethnicrBlack                                -0.18552    0.04764  -3.894
## ethnicrHispanic                             -0.42890    0.04696  -9.134
## ethnicrOther race                           -0.29057    0.06287  -4.622
## ethnicr2 or more races                       0.28139    0.10977   2.564
## OUTSIDE_PAST_YEARRPrior to past year ID theft  0.50006    0.03242  15.424
## notify_breachrData breach victim             0.61300    0.03278  18.703
##                                             Pr(>|z|)
## (Intercept)                                  < 2e-16 ***
## incomer$25,000-$49,999                      6.78e-05 ***
## incomer$50,000-$74,999                       < 2e-16 ***
## incomer$75,000 or more                       < 2e-16 ***
## agerAge 18 to 24                            1.54e-14 ***
## agerAge 25 to 34                             < 2e-16 ***
## agerAge 35 to 49                             < 2e-16 ***
## agerAge 50 to 64                             < 2e-16 ***
## agerAge 65 or older                         5.93e-15 ***
## ethnicrBlack                                9.85e-05 ***
## ethnicrHispanic                              < 2e-16 ***
## ethnicrOther race                           3.81e-06 ***
## ethnicr2 or more races                        0.0104 *
```

```
## 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: 45606  on 66861  degrees of freedom
## Residual deviance: 43574  on 66847  degrees of freedom
## AIC: 43604
##
## Number of Fisher Scoring iterations: 6
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 16663  1316
##          1  8914  1761
##
##                Accuracy : 0.643
##                  95% CI : (0.6374, 0.6485)
##     No Information Rate : 0.8926
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1073
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6515
##             Specificity : 0.5723
##          Pos Pred Value : 0.9268
##          Neg Pred Value : 0.1650
##              Prevalence : 0.8926
##          Detection Rate : 0.5815
##    Detection Prevalence : 0.6275
##       Balanced Accuracy : 0.6119
##
##        'Positive' Class : 0
##
```

```
## [1] "The recall of the above model is 0.926803492964014."
```

```
## [1] "The precision of the above model is 0.651483754936075."
```

```
## [1] "The F1 score of the above model is 0.765129947653595."
```

In an attempt to improve the model used to predict past year identity theft, changes to the data itself were made. Since the data is imbalanced in terms of the outcome variable (89% with no past year identity theft vs 11% reporting identity theft in the past year), random undersampling of the cases where there was no past year identity theft, and keeping all of the cases in which the respondent reported identity theft in place. This resulted in a dataset with an outcome variable that had the same number of cases in which the respondent reported identity theft and respondents reported no past year identity theft (10,258). While the precision remained the same as the previous model, the accuracy score went down (.61), the F1 score (.62), recall (.60), and precision (.64) all decreased compared to the previous model, suggesting that undersampling or some other adjustment to the outcome variable might not be necessary for ITS data with its imbalanced outcome. Therefore, the second logistic regression model is suggested to predict past year identity theft.

```
##
## No past year ID theft    Past year ID theft
##                 10258                 10258


##
## Call:
## glm(formula = idtheft ~ incomer + ager + ethnicr + OUTSIDE_PAST_YEARR +
##     notify_breachr, family = "binomial", data = trainSet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9121  -1.1097   0.1396   1.0967   2.4169
##
## Coefficients:
##                                            Estimate Std. Error z value
## (Intercept)                                -2.55582    0.26827  -9.527
## incomer$25,000-$49,999                      0.23763    0.05537   4.292
## incomer$50,000-$74,999                      0.43161    0.05864   7.360
## incomer$75,000 or more                      0.69758    0.05197  13.423
## agerAge 18 to 24                            1.80181    0.27256   6.611
## agerAge 25 to 34                            2.15672    0.26803   8.047
## agerAge 35 to 49                            2.20291    0.26686   8.255
## agerAge 50 to 64                            2.05704    0.26669   7.713
## agerAge 65 or older                         1.85245    0.26757   6.923
## ethnicrBlack                               -0.27440    0.06145  -4.465
## ethnicrHispanic                            -0.43704    0.05980  -7.308
## ethnicrOther race                          -0.30951    0.08353  -3.705
## ethnicr2 or more races                      0.13565    0.16124   0.841
## OUTSIDE_PAST_YEARRPrior to past year ID theft  0.46339    0.04845   9.564
## notify_breachrData breach victim            0.70922    0.05053  14.035
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## incomer$25,000-$49,999                     1.77e-05 ***
## incomer$50,000-$74,999                     1.84e-13 ***
## incomer$75,000 or more                      < 2e-16 ***
## agerAge 18 to 24                           3.82e-11 ***
## agerAge 25 to 34                           8.51e-16 ***
## agerAge 35 to 49                            < 2e-16 ***
## agerAge 50 to 64                           1.23e-14 ***
## agerAge 65 or older                        4.42e-12 ***
## ethnicrBlack                               8.00e-06 ***
```

```
## ethnicrHispanic                                 2.70e-13 ***
## ethnicrOther race                               0.000211 ***
## ethnicr2 or more races                          0.400182
## 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: 19910  on 14361  degrees of freedom
## Residual deviance: 18761  on 14347  degrees of freedom
## AIC: 18791
##
## Number of Fisher Scoring iterations: 4
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1984 1302
##          1 1093 1775
##
##                Accuracy : 0.6108
##                  95% CI : (0.5985, 0.623)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2216
##
##  Mcnemar's Test P-Value : 2.136e-05
##
##             Sensitivity : 0.6448
##             Specificity : 0.5769
##          Pos Pred Value : 0.6038
##          Neg Pred Value : 0.6189
##              Prevalence : 0.5000
##          Detection Rate : 0.3224
##    Detection Prevalence : 0.5340
##       Balanced Accuracy : 0.6108
##
##        'Positive' Class : 0
##
```

```
## [1] "The recall of the above model is 0.60377358490566."
```

```
## [1] "The precision of the above model is 0.64478388040299."
```

```
## [1] "The F1 score of the above model is 0.623605217664624."
```