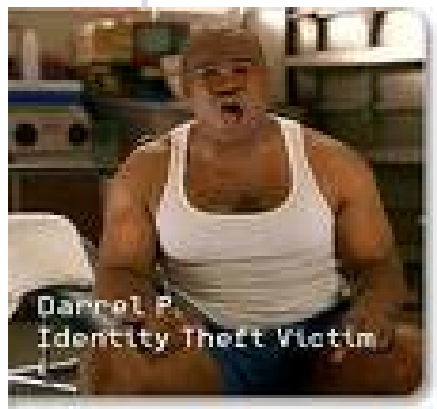
Identity Theft

What does a victim look like?

Mehmet Hondur
Benjama Kounthongkul
Patcharaporn Makarasara
Brenda Martineau
Sophie Shuklin



http://www.youtube.com/watch?v=0cFo7PREzyA

Outline

- Project Goals/Research Questions
- Data Source
- Data variables
- Methodology
  - Exploratory Analysis
  - Data Cleaning
  - Logistic Regression
- Findings
- Recommendations

Identity Theft

- Goal: Understand the characteristics related to being a victim of identity theft
- Research Questions:
  - Are men or women more prone to being victims of identity theft?
  - Are there differences in victimization depending on where you live? Region and urban vs. rural setting
  - Is the minority population more at risk?
  - Does internet use make a difference?
  - Are grocery stores primary places of vulnerability?
  - Does having a higher income mean you will be a victim more often?

Data Source

- Federal Trade Commission, Identity Theft Survey Report, Synovate, September 2003.
  - http://www.ftc.gov/os/2003/09/synovatereport.pdf
- Data sample included 4,057 observations with 46 variables obtained from 4 surveys. 700 experienced identity theft (17.25%)

Sample Data

sample ID	region quota	sex	head of household	primary grocery shopper	education	employment	married	no. of people in HH	home owner	hispanic	race	income	age	internet at home	internet at work	experience theft before	other misuse of personal info
1	4	2	2	2	5	4	1	2	1	2	1	0	2	2		1	2
2	4	2	1	1	4	4	1	3	1	2	1	5	1	1		2	2
3	4	1	1	2	6	1	1	2	1	2	1	9	4	1	1	2	2
4	4	2	1	1	5	4	1	2	9	9	9	0	2	1		2	2
5	4	2	1	1	6	2	1	2	1	2	1	7	5	1	1	2	2
6	4	1	1	1	6	1	2	1	1	2	1	8	5	1	1	2	2
7	4	2	1	1	5	1	1	2	1	1	4	5	1	1	1	2	1
8	4	2	1	1	5	3	1	2	1	2	1	4	6	1		2	1
9	4	1	1	1	5	1	1	3	1	2	4	9	4	1	1	2	2
10	4	1	1	2	3	1	1	2	1	2	1	0	6	1	1	2	2

Variables

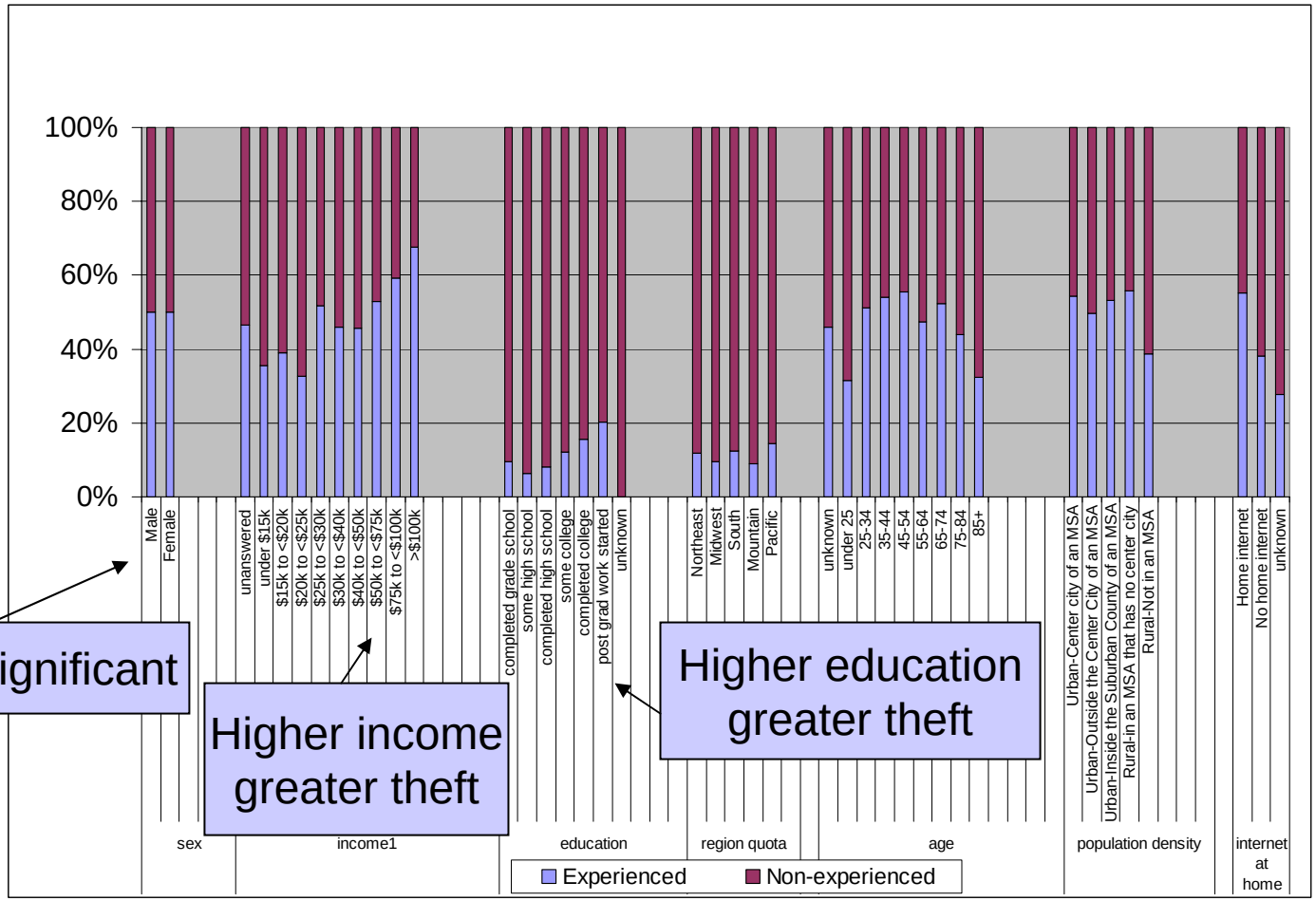
Response Variable: Combined ID Theft = 1

- Credit card misuse
- Other existing accounts misuse
- Other misuse of personal information

Explanatory Variables (46 total)

- Age, gender, race, married, education level
- Income, head of household, primary grocery shopper, # people in household

Initial Exploration



Gender insignificant

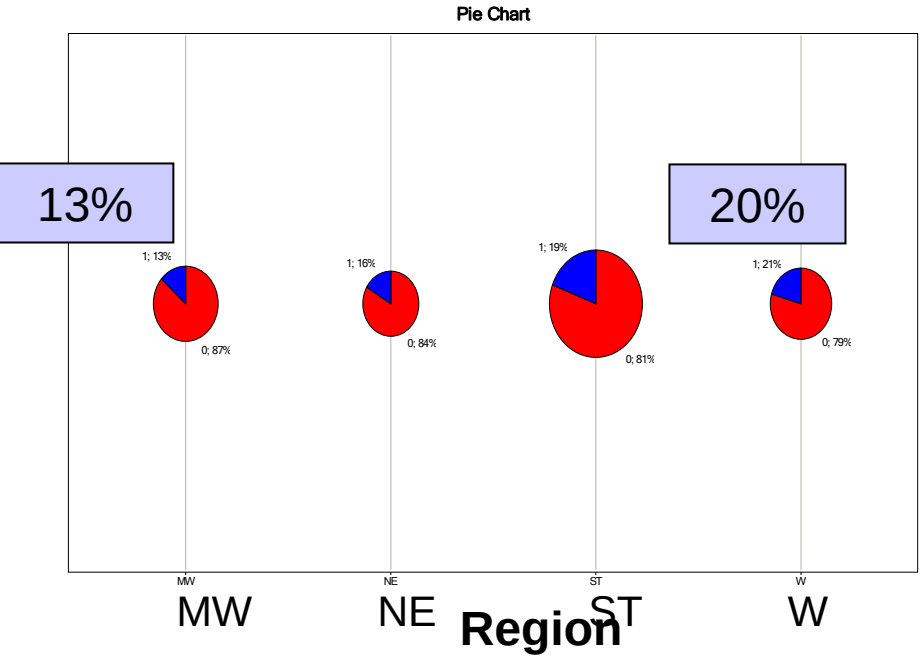
Higher income greater theft

Higher education greater theft

Data Cleaning

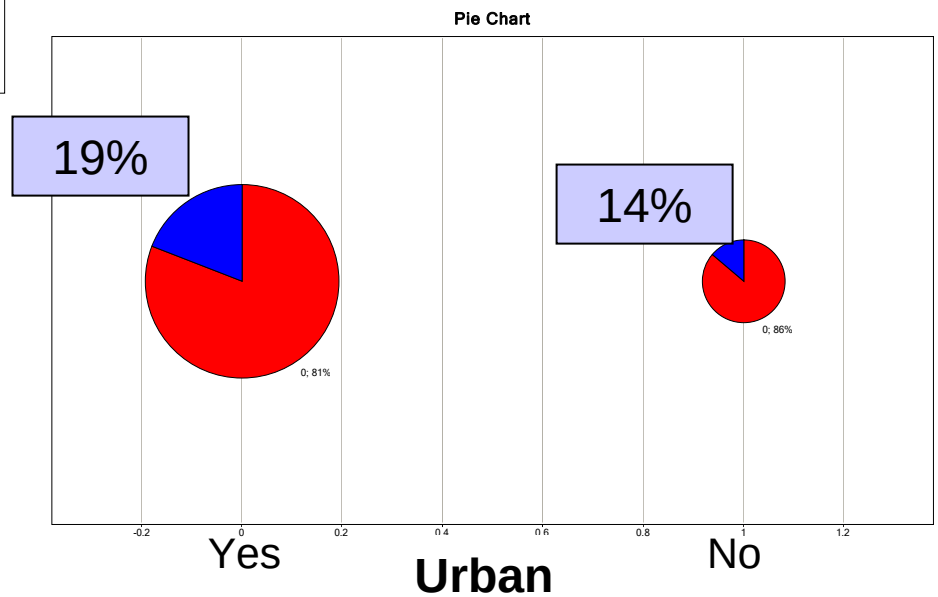
- **Eliminated obvious duplication**
  - Age, regional variables
- **Managed missing data**
  - Deleted observations with missing values
  - Deleted uncertain and unrealistic values for each variable
  - Imputed missing age values using average age (120 records)
  - Imputed income using K Nearest Neighbor (KNN) for income in 9 bins using midpoint (651 records)

Exploratory Analysis

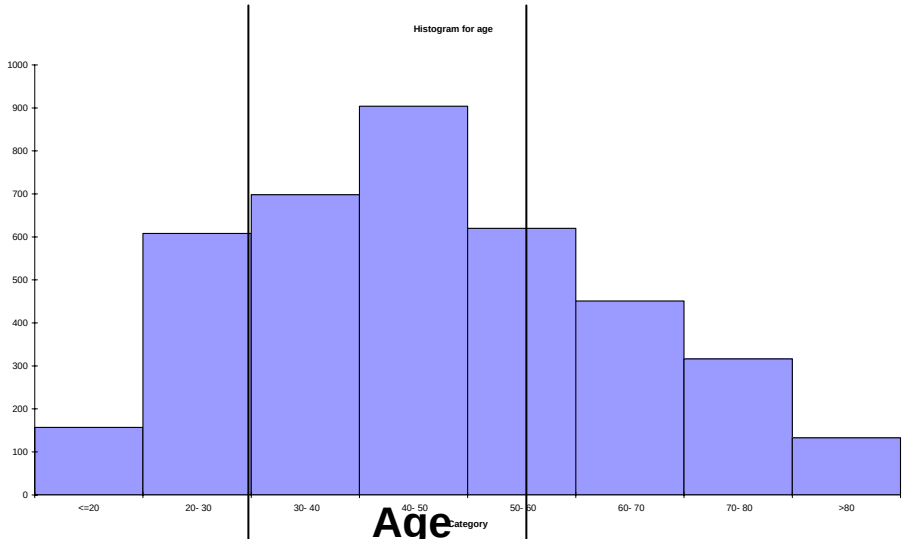


People who live in West region are most prone to identity theft

People who live in urban areas are more prone to identity theft



Exploratory Analysis

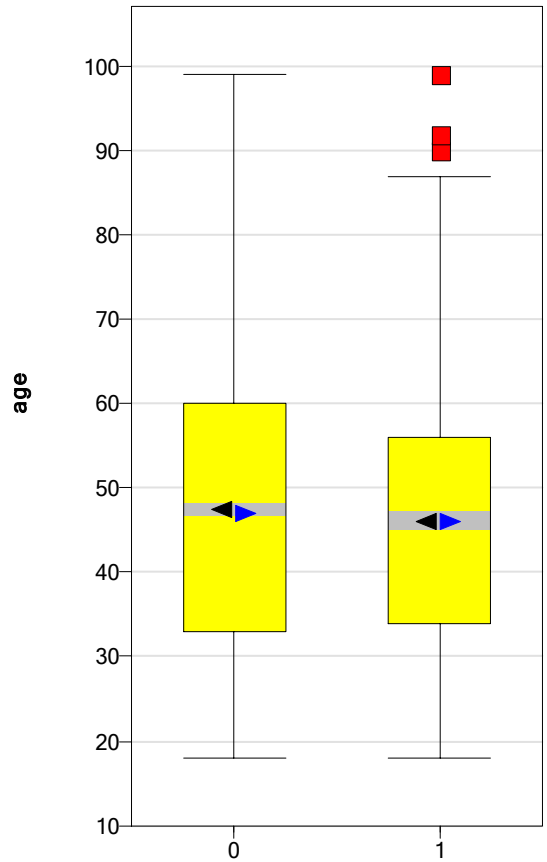


Age

Young
0-30

Middle
31-55

Old
>55



	0	1
Range	81.0	81.0
Mean	47.5	46.1
UAV	99.0	87.0
LAV	18.0	18.0

Combined ID theft

Variables after Imputation

Numerical Variable (2)

- *Income midpoint (k):* The median income of the income group the respondent belongs to
- *Number of People in Household*: The number of people living in the household of the respondent

Categorical Variables (13)

- *Income with missing data binning*
- *Rural*
- *Gender*
- *Head of Household*
- *Primary Grocery Shopper*
- *Age*
- *High Education*
- *Home owner*
- *Race*
- *Married*
- *Employment*
- *Region*
- *Combined Internet*

Findings – Best Model 1

Success Class =1 Cut off = 0.25

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.721699	0.18792033	0	*
rural	-0.2861056	0.10188214	0.0049819	0.75118327
head of HH	0.35316476	0.16072637	0.02799872	1.42356563
High education	0.36140341	0.09561712	0.00015702	1.43534231
region_ST	0.33606958	0.09847201	0.00064289	1.39943635
region_West	0.37880364	0.11104752	0.00064681	1.46053624
combined internet	0.39212537	0.11504573	0.00065338	1.48012328
income with missing data binning_high	0.2267748	0.09332977	0.01510621	1.25454736
age_Bin_Middle	0.22234415	0.08875898	0.012244	1.24900115

Residual df	3878
Residual Dev.	3502.971191
% Success in training data	17.5971186
# Iterations used	8
Multiple R-squared	0.03144305

Training Data scoring - Summary Report

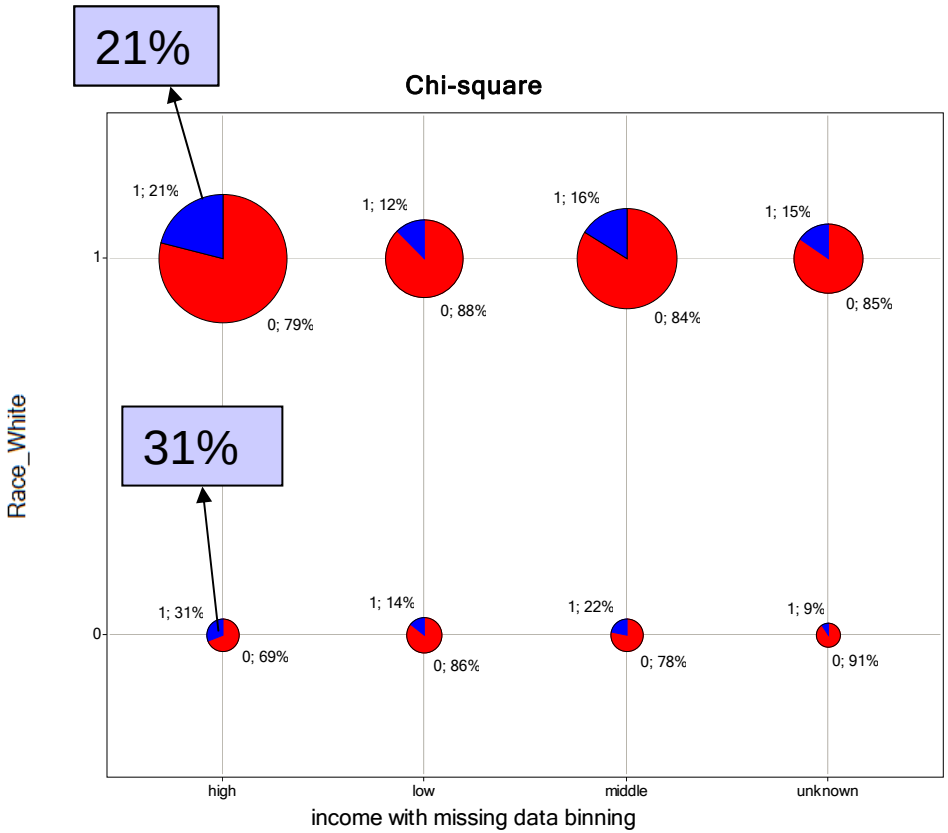
Cut off Prob.Val. for Success (Updatable)	0.25

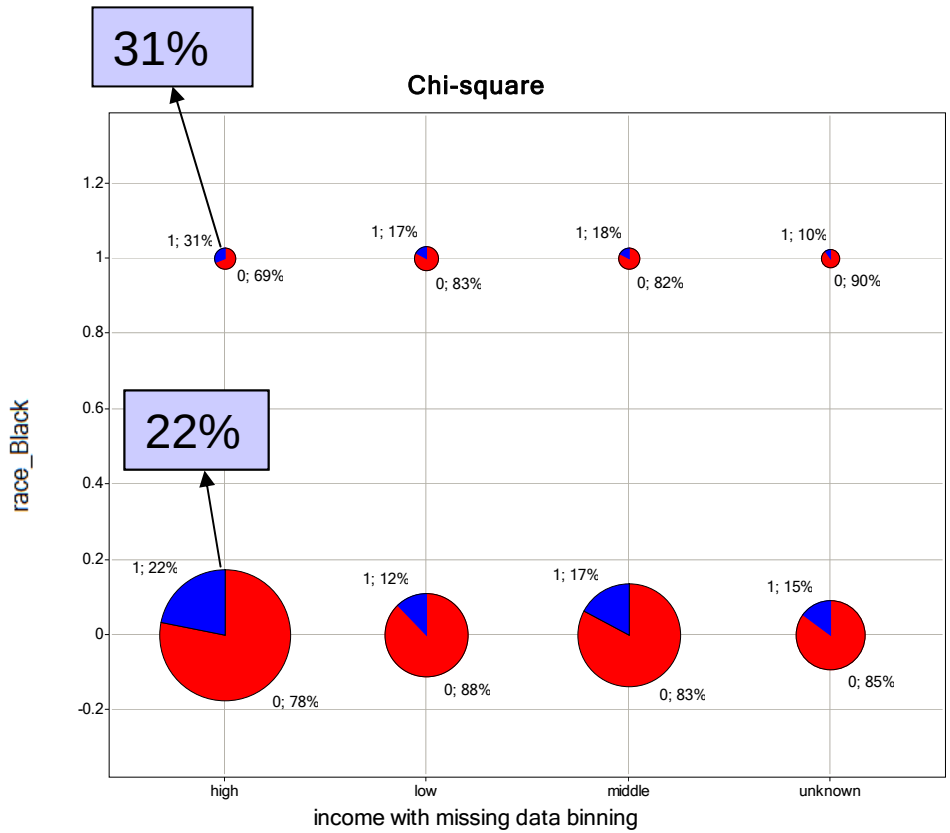
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	168	516
0	440	2763

Error Report			
Class	# Cases	# Errors	% Error
1	684	516	75.44
0	3203	440	13.74
Overall	3887	956	24.59

Why doesn't race have impact?

Binned Income vs. Race





Findings – Best Model 2

The Regression Model Success Class =1 Cut off = 0.25

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.88297915	0.19261804	0	*
rural	-0.2596491	0.10230482	0.01114896	0.77132219
head of HH	0.3963179	0.16219139	0.01454476	1.48634171
High education	0.32000002	0.09767967	0.00105283	1.37712777
income midpoint (k)	0.01073969	0.00227312	0.00000231	1.01079762
region_ST	0.32555714	0.09886307	0.00099121	1.38480198
region_West	0.35945141	0.1115823	0.00127565	1.43254328
combined internet	0.31102306	0.11866625	0.00876748	1.36482072
age_Bin_Middle	0.17830224	0.09010424	0.04783356	1.1951865
income midpoint*race_white	-0.00535986	0.00186799	0.00411359	0.99465448

Residual df	3877
Residual Dev.	3486.922119
% Success in training data	17.5971186
# Iterations used	9
Multiple R-squared	0.03588055

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.25

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	171	513
0	391	2812

Error Report			
Class	# Cases	# Errors	% Error
1	684	513	75.00
0	3203	391	12.21
Overall	3887	904	23.26

Interpretation

From both models:

- ***People in urban areas*** are more prone to identity theft compared to those in rural areas
- ***Heads of household (HH)*** tend to be targets of identity theft more than those who are not HH.
- ***People with higher education*** are more prone to identity theft than those with lower education
- ***People who live in the South and West region*** are more prone to identity theft than those who live in the Midwest
- ***People who have internet*** either at home or at work are involved in identity theft more than those who do not have internet

Interpretation - continued

- ***People who are between 31-55 years old*** are more prone to identity theft than those who are younger
- ***High income people*** tends to be a target for identity theft more than those with low income

From model 2:

- White people are less prone to identity theft than people of other races regardless of the level of their income

Recommendations

- Structure the survey better
- Do further study on why Western and Southern regions might be more vulnerable to ID theft
- Improve surveys with additional questions:
  - How do you dispose of personal papers?
  - Do you use software encrypted sites when on-line?
  - Do you pass personal information via wireless?
- Examine why those who don’t answer income questions are less likely to be victimized