Team Project

Characterizing victims of identity theft

By

Mehmet Hondur
Benjama Kounthongkul
Brenda Martineau
Patcharaporn Makarasara
Sophie Shuklin

BUDT 733
Prof. Galit Shmueli
May 10, 2007

Executive Summary

Background and Task

We acquired an Identity Theft Survey report from the Federal Trade Commission administered in September 2003 by Synovate Research Company[1]. The data sample included 4,057 observations with 46 variables obtained from four waves of surveys. Among 4,057 observations, about 700 experienced identity theft (17.25%). Since 60% of our group members personally experienced identity theft before, we were interested in two main things: 1. verify our assumptions of what type of individuals are more prone to be identity theft victims, 2. verify the reported results of the FTC report. Some of the questions we wanted to explore were: 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? Does internet use make a difference?

Data

Since this data comes from a survey, there were many records that had "bad" data. We ended up eliminating about 4% of the data due to missing, invalid or inconsistent occurrences. As a result we went from 4,057 records down to 3,887 with 17.59% victims.

Analysis Results

From our findings we concluded that the following profile of individual is more prone to identity theft – those that live in an urban area, are head of household, have high education, live in South or West regions as opposed to the Midwest, have Internet at home or work, are between the ages of 31 and 55, are high income and whites are more likely to be victims regardless of how much they earn. We also saw that those who do not report income have a lower incidence of theft.

Recommendations

Based on our analysis and findings we were able to isolate the victims of identity theft to three primary characteristics: high education, West and South region and high income. From the consumer point of view those who attend college and beyond need to ensure they use their personal information carefully. Also, those individuals that live in the South or West should be more discrete. Those that earn high income also need to ensure they know how to protect themselves. Federal and local governments can use this information to pursue training for those in law enforcement as well as those at risk. Businesses should be aware of these findings and use them to educate employees and customers. These actions could reduce societal costs of identity theft which was at $47.6 billion in 2002, according to the FTC report.

Survey administration – there are number of ways to improve the administration of the survey by 1) organizing questions that are not repetitive 2) asking more concrete questions and ensuring that the same questions are being asked across all rounds. Some questions that might help are: how do you dispose of personal papers? Do you use software encrypted sites when on-line? Do you pass personal information via wireless communications?

[1] http://www.ftc.gov/os/2003/09/synovatereport.pdf

Technical Analysis

Exploration, cleaning and data processing

Our goal in this project is explanatory. We want to understand what the various characteristics of identity theft victims are. In order to better understand the data, we decided to visualize it first with bar charts comparing single variables against the response variable (Combined ID theft 1 = victim, 0 = non-victim). We noted gender was not significant; however, higher income and education showed greater incidence of theft. Since most of our data is categorical, we then created pie charts to see the proportion of those who were victims of identity theft vs. those who were not and then again compared across the categories of each predictor to get an idea of what variables have power in explaining the response variable. For example, from the Combined Internet (Internet at Work and Home) variable (see Exhibit 1) we observed that people who have Internet at home and work had 20% of victims whereas those that did not have Combined Internet had 11% victims. So we decided to keep this variable as a significant one in separating the two groups. We then started taking out duplicate variables and records with unrealistic values. We finally imputed the data on age variable using the average value and imputed data in the 'unknown category' by using the KNN method with mid-point income. Since the error rate for this imputation is pretty high (overall 75%), we decided to create another variable with income binning as high, middle, low and unknown so that we do not ignore this category.

We then visualized the data again to reduce variables by binning the categories in a way that can facilitate answering our questions. For example, when we looked at the box-plot of age, we saw that the range of age of non-victims is bigger than that of victim but the average is a little different among two groups so we believed a breakdown of age would help explain better. We then used histogram to frame the range of age to young, middle and old. Next, we created dummy variables as needed and used visualization to determine the reference category e.g. when we look at the % of victim under each race we found that race_White =17%, race_Black =20%, race_Asian = 20% and race_other = 21% so we used race_White as the reference category and we were left with the variables as listed in the definition page.

Discriminant Analysis

We wanted to know the order of importance of the variables in explaining the response variable. Therefore, we proceeded to run the Discriminant Analysis at cutoff 0.6 with Combined ID theft =1 as success class and normalize numeric data for scale 0-1 because we had a lot of categorical predictors. Then we tried to find the variables that are useful in distinguishing between victim and non-victim by comparing the difference of the classification scores and rank the ones with the difference higher than the absolute value of 0.2. The remaining predictors in order of importance were: income midpoint (standardized), region_West, head of household, region_ST, race_other, high education, race_black, combined internet and rural. In this case we used mid-point income with an overall error rate of 25.39%. The rank is region_west, region_south, high education, race_other, head of household, income bin_high, combined internet, race_black, age_bin_middle, rural and employment_retire. In the second case we used bin-income and bin age with overall error rate of 25.21%. However, we did not use this model as our final model because it does not help us in giving meaningful interpretation.

Logistic regression (LR)

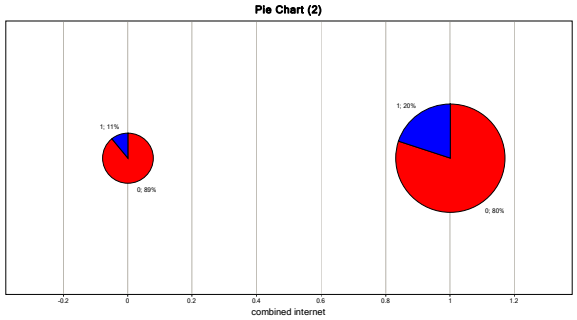
We ran logistic regression to find a model with meaningful interpretation. We ran the model and eliminated variables one by one based on p-values at the 5% significance level and found the best model attached as Exhibit 3. The remaining predictors are rural, head of household, high_education, region_south, region_west, combined_Internet, Income_bin_high and age_bin_middle. The overall error rate of the model is 24.59% and multiple R^2 is .0314. We also ran LR with the exhaustive search option which confirmed our findings that the final model predictors were the ones we should consider (See Exhibit 4).

However, with our knowledge from discriminant analysis, we found that race should have an impact on the model but it was eliminated in LR so we were curious that something might have an impact on race to reduce the power of it. Therefore, we made pie charts showing multiple variable comparisons (Exhibit 5) and we observed that if the individual is Race_Black they are much more prone to identity theft among all income categories. The same pattern is found among race Asian and Other. However, when the individual is Race_White they are less prone to theft among all the income categories. From these findings, we ran a new model with interaction term on Income_Midpoint*Race_White because such characteristics apply to all income categories and found that it was significant. The variables left at the 5% significance level are rural, head of household, high education, income midpoint, region_ST, region_West, combined internet, age_bin_Middle and income midpoint*race_White. The overall error rate of the model is 23.25% and multiple R^2 is .0358, which is better than the first one.

From both models we observe the following:

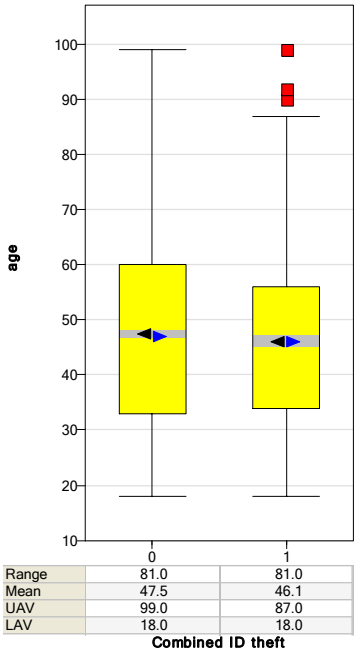
- People in urban areas are more prone to identity theft compared to those in rural areas. More specifically for individuals in urban areas as opposed to rural, the odds of being victims of identity theft are 1.3312 for model 1 and 1.2964 for model 2 keeping other factors in each model the same.
- Head of household (HH) tends to be a target for identity theft more than those who are not i.e. the odds of being victims of identity theft for individuals who are head of household are 1.4236 for model 1 and 1.4863 for model 2 keeping other factors in each model the same.
- People who live in the South and West regions are more prone to identity theft than those who live in the Midwest. For those who live in the West, the odds of being identity theft victims are 1.4605 for model 1 and 1.4325 for model 2 than those in the Midwest keeping other factors the same.
- People with high education i.e. some degree or higher are more prone to identity theft than those with lower education.
- People who have internet either at home or at work are involved in identity theft more than those who do not have internet.
- People who are between 31-55 years old are more prone to identity theft than those who are younger or older.
- High income people (those who earn $50,000 or above) tend to be targets for identity theft more than those with low income.
- From the second model we also see that white people are less prone to identity theft than people of other races regardless of the level of income.

Exhibit 1 – example of pie chart (combine Internet variable)



Red – Non victims
Blue – Victims of identity theft

Exhibit 2 – box plot of age variable



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

Exhibit 3 – Model 1

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

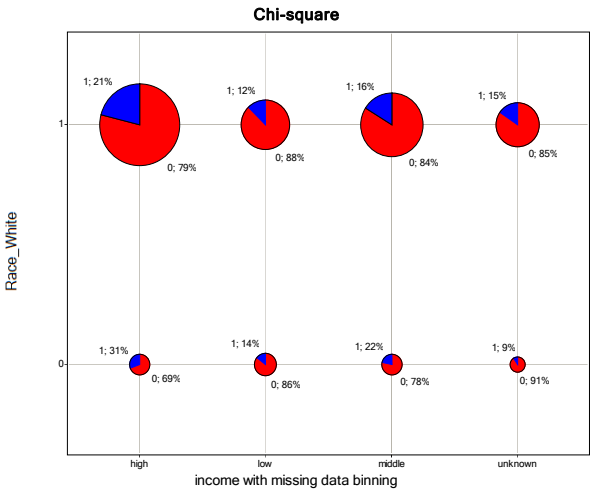
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	168	516
0	440	2763

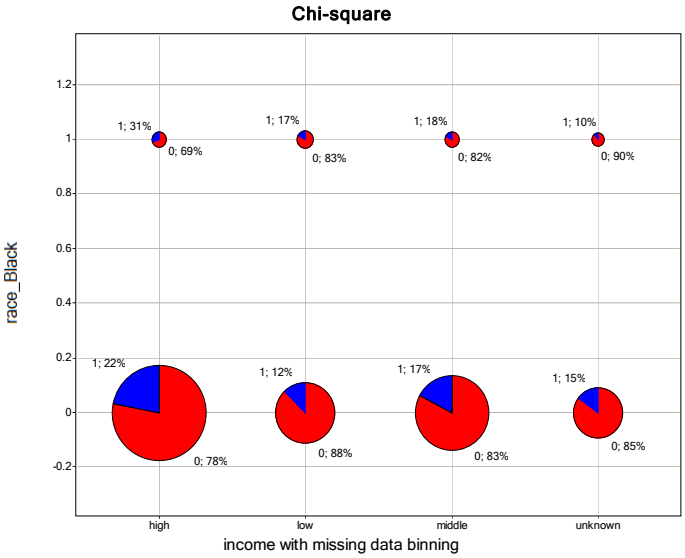
Error Report			
Class	# Cases	# Errors	% Error
1	684	516	75.44
0	3203	440	13.74
Overall	3887	956	24.59

Exhibit 4 – Exhaustive search

#Coeffs	RSS	Cp	Probability	Model (Constant present in all models)											
				1	2	3	4	5	6	7	8	9	10	11	12
2	3942.123535	60.14414597	0.00000003	Constant	High education	*	*	*	*	*	*	*	*	*	*
3	3921.002686	41.01782608	0.00002276	Constant	High education	mbined internet	*	*	*	*	*	*	*	*	*
4	3912.311768	34.32466125	0.00022865	Constant	High education	mbined internet	ge_Bin_Middle	*	*	*	*	*	*	*	*
5	3903.959229	27.96995735	0.00191878	Constant	rural	High education	mbined internet	ge_Bin_Middle	*	*	*	*	*	*	*
6	3897.391846	23.40087509	0.00870676	Constant	High education	region_ST	region_West	mbined internet	ge_Bin_Middle	*	*	*	*	*	*
7	3889.130127	17.1370163	0.05653239	Constant	rural	High education	region_ST	region_West	mbined internet	ge_Bin_Middle	*	*	*	*	*
8	3883.302734	13.30811596	0.16722041	Constant	rural	High education	region_ST	region_West	mbined internet	a binning_high	ge_Bin_Middle	*	*	*	*
9	3878.590576	10.59473801	0.34391272	Constant	rural	head of HH	High education	region_ST	region_West	mbined internet	a binning_high	ge_Bin_Middle	*	*	*
10	3875.092529	9.09578514	0.51956803	Constant	rural	head of HH	High education	race_Other	region_ST	region_West	mbined internet	a binning_high	ge_Bin_Middle	*	*
11	3871.366211	7.36850214	0.74536771	Constant	rural	head of HH	High education	race_Black	race_Other	region_ST	region_West	mbined internet	a binning_high	ge_Bin_Middle	*
12	3869.122803	7.12451315	0.84976041	Constant	rural	head of HH	High education	race_Black	race_Other	region_NE	region_ST	region_West	mbined internet	a binning_high	ge_Bin_Middle
13	3867.59375	7.59506464	0.89910883	Constant	rural	head of HH	High education	race_Black	race_Other	region_NE	region_ST	region_West	mbined internet	a binning_high	binning_middle
14	3866.160156	8.16109943	0.9394502	Constant	rural	head of HH	High education	oyment_Retire	home owner	race_Black	race_Other	region_NE	region_ST	region_West	mbined internet

Exhibit 5 – Pie chart with relationship with income bin





Model 2 – The best model with interaction term of race_white and income midpoint

The Regression Model

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