Classifying Identity Theft based on Victim Profiles
Final Report

Anne Parker
aparker3@stanford.edu

Xueqian Jiang
xqjiang@stanford.edu

Usha Prabhu
uprabhu@stanford.edu

1. Introduction

Identity theft (IdT) occurs when a victim's personally identifying information (PII) is fraudulently obtained and used by another to acquire products and services. Latest estimates from the US Department of Justice indicate that 7% of US households experienced IDT in 2010 up from 5% in 2008. Financial loss estimates from IdT have increased from 13.2 Billion dollars in 2010 to 21 Billion dollars in 2013 (Javelin 2013.) This illustrates both the prevalence of IdT, and the increased economic loss associated with the crime.

A successful IdT crime requires a victim whose identity is stolen and a thief who uses the identity to acquire goods and services via transactions from businesses and government agencies. IdT is multi-faceted, featuring many variations including fraudulent use of existing financial accounts, opening new accounts to acquire products and services, obtaining medical care and services, acquiring government benefits through filing false tax refunds or social security claims, using another's identity for work, obtaining housing, and even as proof of identity for law enforcement officials when charged with a crime. Our focus is on the victims of IdT. IdT impacts individuals in many ways including direct financial loss, emotional distress, and the burden of resolving problems related to the theft of PII, such as repairing credit.

2. Related Work

Considerable research has been devoted to successfully predicting fraudulent transactions, however research focused on understanding IdT victims is largely anecdotal. Much of the existing data consists of self-reported IdT complaints maintained by government and consumer advocacy groups (White, 2008). Persistent lack of data around victims of IdT has resulted in an inability to develop ways of systematically targeting and preventing identity theft through understanding the victim experience. In 2008 the US Department of Justice sponsored a detailed random survey designed to collect information specifically around IdT including information on how victims discovered their IdT, their financial and emotional losses, interactions with law enforcement and other government agencies, and the actions and burden required to resolve their IdT. The data from this survey was released in 2012 and summarized in a special report released by the Bureau of Justice Statistics (Langdon, 2010). Our contribution to the literature around IdT victimization is to holistically model the victim experience using this comprehensive database.

3. Data Overview

Our data source is the 2008 Identity Theft Supplement to the National Crime Victimization Survey (NCVS) produced by the US Census for the Bureau of Justice Statistics. The data comprise a statistically valid sample of the US population with 56,480 individuals from 34,799 households and includes over 250 survey questions designed to measure the extent of identity theft and the economic and emotional costs for individual victims. Although identity theft is a rapidly growing crime it still impacts only a small portion of the population. Of the 56,480 individuals in the survey, only 2,815 experienced an attempted or successful IdT incident within the two years prior to 2008. Our project includes only these 2,815 victims.

We categorized these victims by types of IdT using the NCVS standard classification from the IdT Supplement – Credit Card (CC); Other Existing Accounts including bank, debit card loan (EA); Opening New Accounts (NA), together with fraudulent use of personally identifiable information (PII) to obtain employment, housing, medical care, government benefits, and to avoid criminal charges. However we discovered that a number of incidents involved multiple types of IdT so we added an additional category to cover this situation. Sample sizes for each IdT category are provided in Table 1.

# CC	# EA	# NA and PII	# Multiple
1,211	819	352	433

Table 1. Final Project Sample Sizes

We encountered several significant data issues related to the complexity of the survey questionnaire. Questions related to the type of IdT covered both attempted and successful incidents. However depending on the victim's responses, different sets of questions were asked for attempted versus successful IdT incidents. As our categorization included both attempted and successful IdT incidents we merged the two sets of questions. In many cases this required referencing the survey instrument, interviewer training materials, and cognitive survey pre-testing reports to fully understand the data and make sensible choices for the merged questions (DeMaio, 2008.)

Almost all of our features are categorical, and the few that are not we discretized. Although we have few truly missing values for our features, we have many "unknown", "don't know", and "refused" responses to some questions, which were merged together into a single "unknown" category. After data pre-processing, our data set included 88 potential features.

4. **Problem Formulation**

Our hypothesis is that different types of IdT impacts individual segments of the population in different ways. Our goal is to identify and describe the characteristics of these different population segments. We used a two-prong approach to meet this goal. Using the IdT categorizations specified by the authors of our data study, we used supervised machine learning techniques to understand the characteristics of each category. Simultaneously, we used unsupervised machine learning techniques to identify patterns within the data that might suggest alternate IdT categorizations.

Our supervised ML methodology consisted of three steps: feature selection, prediction and rule learning. Feature selection was done using SAMME, a boosting ensemble method extended to the multi-class problem. We used a CART based implementation available in the R package adabag. We used the features selected by SAMME to predict IdT type by using an SVM model with a polynomial kernel of degree 3 with L1 regularization. We then used SimpleCART to understand the rules behind these predicted categories.

Independently, we used CascadeSimpleKMeans (a Weka package) to cluster our data. We then used a C4.5 decision tree to understand the rules behind the clusters.

5. **Results and analysis**

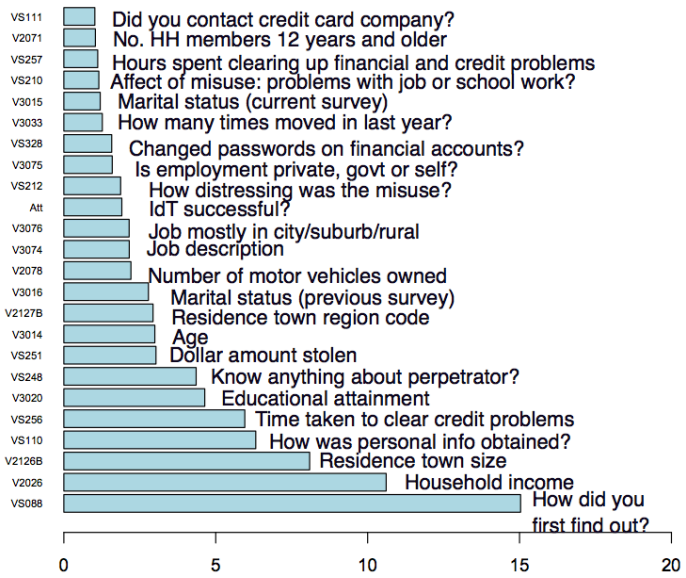

Figure 1. Feature selection using SAMME. X-axis shows GINI index gain.

Based on the SAMME results, we eliminated 22 features with Gini Index gain equal to zero leaving 66 features used in our analysis. The top 24 features are displayed in Figure 1.

To predict IdT type we used an SVM model using a polynomial kernel of degree 3 with L1 regularization. The model was tuned using 10-fold cross validation on the training set together with a grid search on the model hyper-parameters with a best result of gamma = 0.001, offset = 1, cost = 10. This SVM algorithm uses a one-against-one together with a voting scheme for multi-class prediction.

We achieved a test accuracy rate of 69% and a training accuracy rate of 72%. SVM model learning curves are provided in Fig. 2. The test accuracy continues to dip with increase in training size, suggesting that more data may help increase our training accuracy.

	Actual IdT Type			
Predicted IdT Type	# CC	# EA	# NA and PII	# Multiple
# CC	1,099	324	115	93
# EA	84	471	54	54
# NA and PII	28	24	183	16
# Multiple	0	0	0	270

Table 2. SVM Results – Confusion Matrix

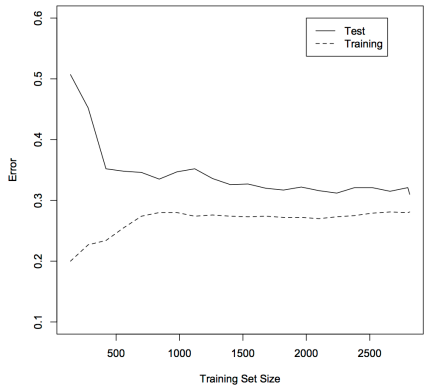

Figure. 2: SVM Learning Curves

The strongest SAMME predictor is VS088 – How Did You First Find Out About Misuse? The categories for this feature include items that generally map back directly to the types of IdT making this feature an excellent choice for prediction but less useful for understanding the interplay between victim demographics and victim behavior. Thus we decided to use this feature for SVM model prediction but not for rule-learning. We then ran the SimpleCART algorithm. The rule for obtaining the tree branches is with a cutoff of at least 20% population of that class in the leaf node. The tree for rule learning is shown in Figures 3 and 4. Note that all the variables in the tree have high GINI index gain, as shown in Figure 1.

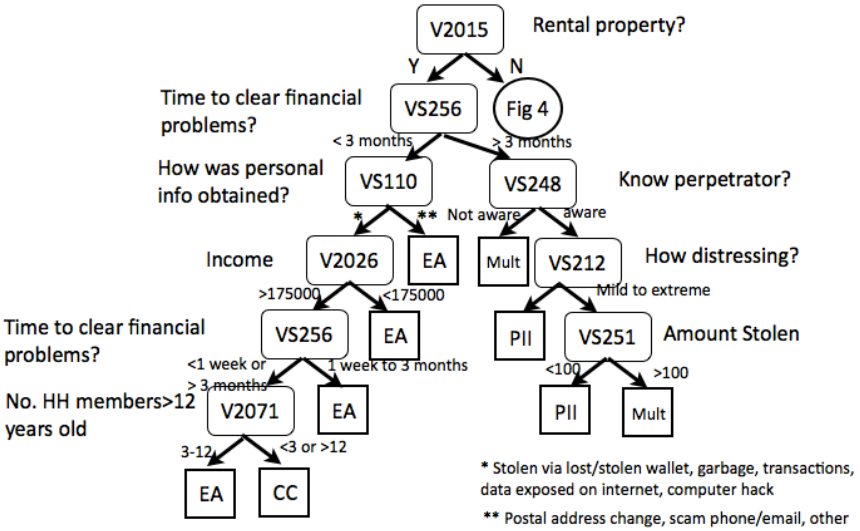

Figure 3: Decision tree created using SimpleCART on predicted results of SVM (part 1)

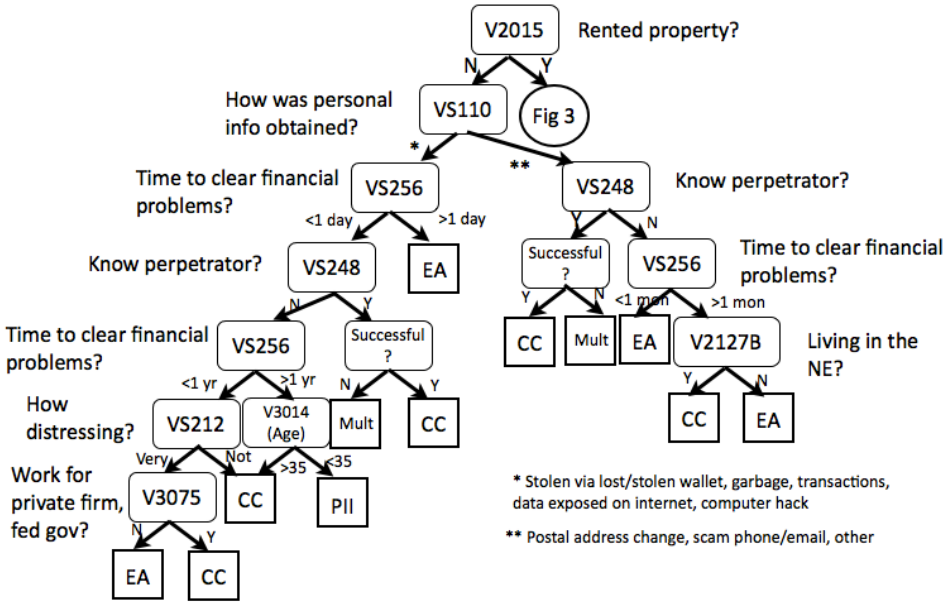

Figure 4: Decision tree created using SimpleCART on predicted results of SVM (part 2)

Here are the inferences we made on each class:

Class 1 (Credit Card Theft):
The time required to resolve financial and credit problems is small, resulting in low levels of victim distress. Victims of this type of IdT tend not to know who stole their Id. They most likely had their ID stolen thru lost/stolen wallet, garbage, data exposed on internet, or computer hacks. It is more prevalent in the north-east. While credit card theft comprises a majority of our data, most credit card theft is not successful.

Class 2 (Other Existing Accounts):
The victims most likely had their ID stolen thru the post office (mail, address change), in responce to internet scam, or stolen from personal files. It is less prevalent in the north east. It takes longer to resolve financial and credit problems. Most victims have lower incomes than class 1.

Class 3 (PII and New Accounts):
This category can have a higher level of emotional distress. The financial losses are usually small. It takes greater than 3 months to resolve financial and credit problems. The victims are generally younger.

Class 4 (Multiple):
These victims tend to know who stole their ID. This category usually includes larger dollar amounts of theft, accompanied by higher levels of emotional distress. These crimes are more often successful, generally taking longer to process resulting financial problems.

In parallel, we ran CascadeSimpleKMeans to explore whether any natural groupings and patterns were present in the data. We found two distinct clusters. We identify these clusters as cluster 0 and cluster 1 and note that each have similar distributions of IdT category types used above. We then used the C4.5 algorithm for insight into the characteristics around each cluster. The results are shown in Figure 5.

The tree presents an interesting story around attempted versus actual IdT. People taking precautions against IdT are less likely to be victims of actual IdT as opposed to attempted IdT. These precautions include using security software, changing passwords, and checking credit reports.

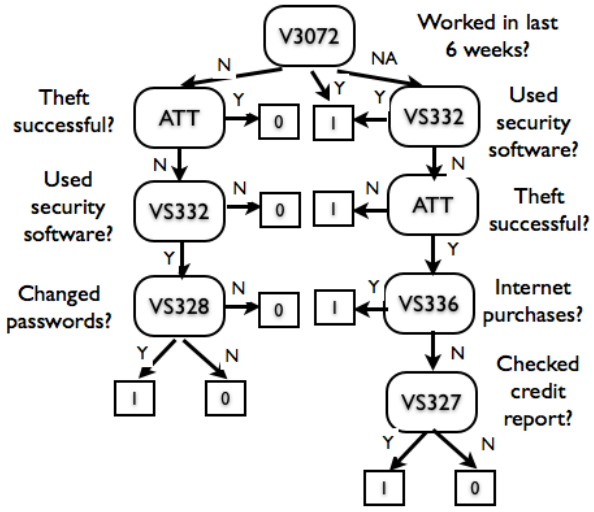

Figure 5: Decision tree created using C4.5 for cluster 0 and cluster 1

6. **Conclusions and Future Work**

Our analysis provides useful insights into IdT victim experiences and does suggest that IdT impacts individual segments of the population in different ways. In addition our clustering results strongly support the value of engaging in precautions designed to prevent IdT.

Our SVM model accuracy rate was lower than we had hoped and was likely impacted by having unbalanced IdT classes. The largest class – credit card IdT had the largest accuracy rate at 91% followed by accuracy rates of 58% for existing accounts, 52% for new accounts and use of PII, and 62% for multiple incidents of IdT. Incorporating methods for handling unbalanced classes may improve accuracy results.

Our data is challenging due to survey respondents difficulty characterizing their IdT experiences into the pre-determined survey IdT categories (DeMaio, 2008) and the large number of "Unknown" answers in the survey responses. Incorporating mechanisms to better handle such unknown values may improve our results.

Finally, we created a new class "Multiple" to deal with survey respondents reporting multiple IdT incidents. Finding a better way to deal with this data may result in cleaner categorizations.

7. **References**

DeMaio, Thersa; Beck, Jennifer, 2008. *"Developing Questionnaire Items to Measure Identity Theft."* Proceedings, American Statistical Association, Survey Research Methods Section, 63rd Annual AAPOR Conference.

Javelin Research and Strategy (2013). Identity theft/fraud statistics. Retrieved from http://www.statisticbrain.com/identity-theft-fraud-statistics

Ji, Zhu, et al. 2009. *"Multi-class Ada Boost."* Statistics and Its Interface Volume 2 pages 349 – 360.

Langton, Lynn; Planty, Michael, "*Victims of Identity Theft, 2008*." Special Report: National Crime Victimization Survey Supplement. NCJ 231680, Washington, DC: United States Department of Justice, Bureau of Justice Statistics, Dec 2010.

National Crime Victimization Survey: Identity Theft Supplement, 2008. United States Department of Justice, Bureau of Justice Statistics Identification Number: 26362

White, Michael; Fisher, Christopher, 2008. "*Assessing Our Knowledge of Identity Theft: The Challenges to Effective Prevention and Control Efforts*." Criminal Justice Policy Review 19(1): 3 - 24