**MS Data Analytics – Research Project 698 – 2/29/2017**

**Credit Fraud Detection and the Introduction of the EMV Chip**

**Max Wagner**

***Introduction***

The United States has recently implemented EMV chips in most common credit and debit cards, comparable the standard set by European countries. EMV chips are said to be an improvement over the traditional strip and signature design. The strips used on prior cards are unchanging, reproducible, and able to be used in multiple transactions. Chips produce a unique and dynamic transaction code at the point of each sale, which makes duplication and reuse significantly more difficult. There are opposing viewpoints that a stronger physical security leads to an increase in “card-not-present” fraud. This can be supported by looking at the introduction of EMV chips in countries like Britain.

***Prior Work***

Fraud detection requires understanding the generalized principles behind it. A paper on data mining techniques by Phua et al. (2005) summarizes different techniques, and what means each of them satisfy. One of their conclusions is that research often focuses extensively on complex algorithms, while simpler, faster methods such as naïve Bayes are ignored. Another problematic area is the unbalanced nature of credit fraud data sources. In later research by the same authors they identify a method of suspicion scoring called CASS (communal analysis suspicion scoring) (Phua et al. 2009). The technique works with a network system that compares credit uses over time and geography. Their preliminary results show that CASS lessens false positives and preserves acceptable true positives.

Often in a database of all credit charges, only a small proportion will be fraudulent, meaning that model formation is difficult. The key problem of undersampling causes a skew in distribution. Pozzolo et al. researched a manner using Bates Minimum Risk theory to find classification thresholds that best fit the data and how to adjust the value over time.

Ngai et al. takes a more overall approach and classifies research on fraud into six different categories of techniques and how each of them have been used in various disciplines of fraud detection. The classification of how frequent and how efficient different types of data mining are is useful to know where to focus, and where to potentially avoid. In particular, logistic models, neural networks, decision trees, and Bayesian networks emerged as the top choices in most situations.

***Problem Formulation, Objectives, & Evaluation***

The goal of the study is primarily to identify the ways in which fraud is detected, how EMV chips affect those ways, and gaining an understanding of how the main methods (logistic models, neural networks, decision trees, and Bayesian networks) coexist to provide a reliable detection model. The United States is in the early stages of widespread use of EMV chips, meaning that other countries’ experiences can be used to identify likely problems and areas of improvement. My secondary aim is to research potential new routes of fraud once chips are in use. Measuring success in the quantitative portion of the project is simple. An algorithm that properly identifies fraudulent charges to a reasonable degree is a success. The qualitative success of the project is more subjective, and can be judged on both the breadth and depth of the information provided to answering the main and secondary objectives.

***Assumptions***

A vital assumption to the success of the study is that there is sufficient public data in order to test the effectiveness of different models, as well as enough modern information on the viability of EMV chips. Potential data sources can be seen below.

***Data Sources***

https://www.ll.mit.edu/ideval/data/

http://www.netresec.com/?page=PcapFiles

http://contagiodump.blogspot.qa/

http://digitalcorpora.org/

http://malware-traffic-analysis.net/

http://www.secrepo.com/

https://isc.sans.edu/api/

https://www.kaggle.com/dalpozz/creditcardfraud

***References***

E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support Systems, Volume 50, Issue 3, February 2011, Pages 559-569, ISSN 0167-9236

Clifton Phua, Ross Gayler, Vincent Lee, Kate Smith-Miles, On the communal analysis suspicion scoring for identity crime in streaming credit applications, European Journal of Operational Research, Volume 195, Issue 2, 1 June 2009, Pages 595-612, ISSN 0377-2217

Clifton Phua, Ross Gayler, Vincent Lee, Kate Smith-Miles. A comprehensive survey of data mining-based fraud detection research. Artificial Intelligence Review. 2005

Andrea Dal Pozzolo, Olivier Caelen, Reid Johnson, Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. 2015 IEEE Symposium Series on Computational Intelligence. 2015