

FRAUD ANALYTICS

USING DESCRIPTIVE,
PREDICTIVE, AND
SOCIAL NETWORK
TECHNIQUES

A GUIDE TO
DATA SCIENCE FOR
FRAUD DETECTION



BART BAESENS
VÉRONIQUE VAN VLASSELAER
WOUTER VERBEKE

WILEY

Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques

Wiley & SAS Business Series

The Wiley & SAS Business Series presents books that help senior-level managers with their critical management decisions.

Titles in the Wiley & SAS Business Series include:

Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications by Bart Baesens

Bank Fraud: Using Technology to Combat Losses by Revathi Subramanian

Big Data Analytics: Turning Big Data into Big Money by Frank Ohlhorst

Big Data, Big Innovation: Enabling Competitive Differentiation through Business Analytics by Evan Stubbs

Business Analytics for Customer Intelligence by Gert Laursen

Business Intelligence Applied: Implementing an Effective Information and Communications Technology Infrastructure by Michael Gendron

Business Intelligence and the Cloud: Strategic Implementation Guide by Michael S. Gendron

Business Transformation: A Roadmap for Maximizing Organizational Insights by Aiman Zeid

Connecting Organizational Silos: Taking Knowledge Flow Management to the Next Level with Social Media by Frank Leistner

Data-Driven Healthcare: How Analytics and BI Are Transforming the Industry by Laura Madsen

Delivering Business Analytics: Practical Guidelines for Best Practice by Evan Stubbs

Demand-Driven Forecasting: A Structured Approach to Forecasting, second edition by Charles Chase

Demand-Driven Inventory Optimization and Replenishment: Creating a More Efficient Supply Chain by Robert A. Davis

Developing Human Capital: Using Analytics to Plan and Optimize Your Learning and Development Investments by Gene Pease, Barbara Beresford, and Lew Walker

The Executive's Guide to Enterprise Social Media Strategy: How Social Networks Are Radically Transforming Your Business by David Thomas and Mike Barlow

Economic and Business Forecasting: Analyzing and Interpreting Econometric Results by John Silvia, Azhar Iqbal, Kaylyn Swankoski, Sarah Watt, and Sam Bullard

Financial Institution Advantage and The Optimization of Information Processing by Sean C. Keenan

Foreign Currency Financial Reporting from Euros to Yen to Yuan: A Guide to Fundamental Concepts and Practical Applications by Robert Rowan

Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data Driven Models by Keith Holdaway

Health Analytics: Gaining the Insights to Transform Health Care by Jason Burke

Heuristics in Analytics: A Practical Perspective of What Influences Our Analytical World by Carlos Andre Reis Pinheiro and Fiona McNeill

Human Capital Analytics: How to Harness the Potential of Your Organization's Greatest Asset by Gene Pease, Boyce Byerly, and Jac Fitz-enz

Implement, Improve and Expand Your Statewide Longitudinal Data System: Creating a Culture of Data in Education by Jamie McQuiggan and Armistead Sapp

Killer Analytics: Top 20 Metrics Missing from Your Balance Sheet by Mark Brown

Predictive Analytics for Human Resources by Jac Fitz-enz and John Mattox II

Predictive Business Analytics: Forward-Looking Capabilities to Improve Business Performance by Lawrence Maisel and Gary Cokins

Retail Analytics: The Secret Weapon by Emmett Cox

Social Network Analysis in Telecommunications by Carlos Andre Reis Pinheiro

Statistical Thinking: Improving Business Performance, second edition by Roger W. Hoerl and Ronald D. Snee

Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics by Bill Franks

Too Big to Ignore: The Business Case for Big Data by Phil Simon

The Value of Business Analytics: Identifying the Path to Profitability by Evan Stubbs

The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions by Phil Simon

Understanding the Predictive Analytics Lifecycle by Al Cordoba

Unleashing Your Inner Leader: An Executive Coach Tells All by Vickie Bevenour

Using Big Data Analytics: Turning Big Data into Big Money by Jared Dean

Win with Advanced Business Analytics: Creating Business Value from Your Data by Jean Paul Isson and Jesse Harriott

For more information on these and other titles in the series, please visit www.wiley.com.

Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques

*A Guide to Data Science
for Fraud Detection*

Bart Baesens
Véronique Van Vlasselaer
Wouter Verbeke

WILEY

Copyright © 2015 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the Web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Baesens, Bart.

Fraud analytics using descriptive, predictive, and social network techniques : a guide to data science for fraud detection / Bart Baesens, Veronique Van Vlasselaer, Wouter Verbeke.

pages cm. — (Wiley & SAS business series)

Includes bibliographical references and index.

ISBN 978-1-119-13312-4 (cloth) — ISBN 978-1-119-14682-7 (epdf) —

ISBN 978-1-119-14683-4 (epub)

1. Fraud—Statistical methods. 2. Fraud—Prevention. 3. Commercial crimes—Prevention. I. Title.

HV6691.B34 2015

364.16' 3015195—dc23

2015017861

Cover Design: Wiley

Cover Image: ©iStock.com/aleksandarvelasevic

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

*To my wonderful wife, Katrien, and kids, Ann-Sophie, Victor,
and Hannelore.*

To my parents and parents-in-law.

*To my husband and soul mate, Niels, for his never-ending
support.*

To my parents, parents-in-law, and siblings-in-law.

To Luit and Titus.

Contents

List of Figures xv

Foreword xxiii

Preface xxv

Acknowledgments xxix

Chapter 1	Fraud: Detection, Prevention, and Analytics!	1
	Introduction	2
	Fraud!	2
	Fraud Detection and Prevention	10
	Big Data for Fraud Detection	15
	Data-Driven Fraud Detection	17
	Fraud-Detection Techniques	19
	Fraud Cycle	22
	The Fraud Analytics Process Model	26
	Fraud Data Scientists	30
	A Fraud Data Scientist Should Have Solid Quantitative Skills	30
	A Fraud Data Scientist Should Be a Good Programmer	31
	A Fraud Data Scientist Should Excel in Communication and Visualization Skills	31
	A Fraud Data Scientist Should Have a Solid Business Understanding	32
	A Fraud Data Scientist Should Be Creative	32
	A Scientific Perspective on Fraud	33
	References	35
Chapter 2	Data Collection, Sampling, and Preprocessing	37
	Introduction	38
	Types of Data Sources	38
	Merging Data Sources	43
	Sampling	45
	Types of Data Elements	46

Visual Data Exploration and Exploratory Statistical Analysis	47
Benford's Law	48
Descriptive Statistics	51
Missing Values	52
Outlier Detection and Treatment	53
Red Flags	57
Standardizing Data	59
Categorization	60
Weights of Evidence Coding	63
Variable Selection	65
Principal Components Analysis	68
RIDITs	72
PRIDIT Analysis	73
Segmentation	74
References	75
Chapter 3 Descriptive Analytics for Fraud Detection	77
Introduction	78
Graphical Outlier Detection Procedures	79
Statistical Outlier Detection Procedures	83
Break-Point Analysis	84
Peer-Group Analysis	85
Association Rule Analysis	87
Clustering	89
Introduction	89
Distance Metrics	90
Hierarchical Clustering	94
Example of Hierarchical Clustering Procedures	97
<i>k</i> -Means Clustering	104
Self-Organizing Maps	109
Clustering with Constraints	111
Evaluating and Interpreting Clustering Solutions	114
One-Class SVMs	117
References	118
Chapter 4 Predictive Analytics for Fraud Detection	121
Introduction	122
Target Definition	123
Linear Regression	125
Logistic Regression	127
Basic Concepts	127
Logistic Regression Properties	129
Building a Logistic Regression Scorecard	131

Variable Selection for Linear and Logistic Regression	133
Decision Trees	136
Basic Concepts	136
Splitting Decision	137
Stopping Decision	140
Decision Tree Properties	141
Regression Trees	142
Using Decision Trees in Fraud Analytics	143
Neural Networks	144
Basic Concepts	144
Weight Learning	147
Opening the Neural Network Black Box	150
Support Vector Machines	155
Linear Programming	155
The Linear Separable Case	156
The Linear Nonseparable Case	159
The Nonlinear SVM Classifier	160
SVMs for Regression	161
Opening the SVM Black Box	163
Ensemble Methods	164
Bagging	164
Boosting	165
Random Forests	166
Evaluating Ensemble Methods	167
Multiclass Classification Techniques	168
Multiclass Logistic Regression	168
Multiclass Decision Trees	170
Multiclass Neural Networks	170
Multiclass Support Vector Machines	171
Evaluating Predictive Models	172
Splitting Up the Data Set	172
Performance Measures for Classification Models	176
Performance Measures for Regression Models	185
Other Performance Measures for Predictive Analytical Models	188
Developing Predictive Models for Skewed Data Sets	189
Varying the Sample Window	190
Undersampling and Oversampling	190
Synthetic Minority Oversampling Technique (SMOTE)	192
Likelihood Approach	194
Adjusting Posterior Probabilities	197
Cost-sensitive Learning	198
Fraud Performance Benchmarks	200
References	201

Chapter 5 Social Network Analysis for Fraud Detection	207
Networks: Form, Components, Characteristics, and Their Applications	209
Social Networks	211
Network Components	214
Network Representation	219
Is Fraud a Social Phenomenon? An Introduction to Homophily	222
Impact of the Neighborhood: Metrics	227
Neighborhood Metrics	228
Centrality Metrics	238
Collective Inference Algorithms	246
Featurization: Summary Overview	254
Community Mining: Finding Groups of Fraudsters	254
Extending the Graph: Toward a Bipartite Representation	266
Multipartite Graphs	269
Case Study: Gotcha!	270
References	277
Chapter 6 Fraud Analytics: Post-Processing	279
Introduction	280
The Analytical Fraud Model Life Cycle	280
Model Representation	281
Traffic Light Indicator Approach	282
Decision Tables	283
Selecting the Sample to Investigate	286
Fraud Alert and Case Management	290
Visual Analytics	296
Backtesting Analytical Fraud Models	302
Introduction	302
Backtesting Data Stability	302
Backtesting Model Stability	305
Backtesting Model Calibration	308
Model Design and Documentation	311
References	312
Chapter 7 Fraud Analytics: A Broader Perspective	313
Introduction	314
Data Quality	314
Data-Quality Issues	314
Data-Quality Programs and Management	315
Privacy	317
The RACI Matrix	318
Accessing Internal Data	319

Label-Based Access Control (LBAC)	324
Accessing External Data	325
Capital Calculation for Fraud Loss	326
Expected and Unexpected Losses	327
Aggregate Loss Distribution	329
Capital Calculation for Fraud Loss Using Monte Carlo Simulation	331
An Economic Perspective on Fraud Analytics	334
Total Cost of Ownership	334
Return on Investment	335
In Versus Outsourcing	337
Modeling Extensions	338
Forecasting	338
Text Analytics	340
The Internet of Things	342
Corporate Fraud Governance	344
References	346

About the Authors 347

Index 349

List of Figures

Figure 1.1	Fraud Triangle	7
Figure 1.2	Fire Incident Claim-Handling Process	13
Figure 1.3	The Fraud Cycle	23
Figure 1.4	Outlier Detection at the Data Item Level	25
Figure 1.5	Outlier Detection at the Data Set Level	25
Figure 1.6	The Fraud Analytics Process Model	26
Figure 1.7	Profile of a Fraud Data Scientist	33
Figure 1.8	Screenshot of Web of Science Statistics for Scientific Publications on Fraud between 1996 and 2014	34
Figure 2.1	Aggregating Normalized Data Tables into a Non-Normalized Data Table	44
Figure 2.2	Pie Charts for Exploratory Data Analysis	49
Figure 2.3	Benford's Law Describing the Frequency Distribution of the First Digit	50
Figure 2.4	Multivariate Outliers	54
Figure 2.5	Histogram for Outlier Detection	54
Figure 2.6	Box Plots for Outlier Detection	55
Figure 2.7	Using the z -Scores for Truncation	57
Figure 2.8	Default Risk Versus Age	60
Figure 2.9	Illustration of Principal Component Analysis in a Two-Dimensional Data Set	68
Figure 3.1	3D Scatter Plot for Detecting Outliers	80
Figure 3.2	OLAP Cube for Fraud Detection	80
Figure 3.3	Example Pivot Table for Credit Card Fraud Detection	82

Figure 3.4	Break-Point Analysis	84
Figure 3.5	Peer-Group Analysis	86
Figure 3.6	Cluster Analysis for Fraud Detection	91
Figure 3.7	Hierarchical Versus Nonhierarchical Clustering Techniques	91
Figure 3.8	Euclidean Versus Manhattan Distance	92
Figure 3.9	Divisive Versus Agglomerative Hierarchical Clustering	94
Figure 3.10	Calculating Distances between Clusters	95
Figure 3.11	Example for Clustering Birds. The Numbers Indicate the Clustering Steps	96
Figure 3.12	Dendrogram for Birds Example. The Thick Black Line Indicates the Optimal Clustering	96
Figure 3.13	Screen Plot for Clustering	97
Figure 3.14	Scatter Plot of Hierarchical Clustering Data	98
Figure 3.15	Output of Hierarchical Clustering Procedures	98
Figure 3.16	<i>k</i> -Means Clustering: Start from Original Data	105
Figure 3.17	<i>k</i> -Means Clustering Iteration 1: Randomly Select Initial Cluster Centroids	105
Figure 3.18	<i>k</i> -Means Clustering Iteration 1: Assign Remaining Observations	106
Figure 3.19	<i>k</i> -Means Iteration Step 2: Recalculate Cluster Centroids	107
Figure 3.20	<i>k</i> -Means Clustering Iteration 2: Reassign Observations	107
Figure 3.21	<i>k</i> -Means Clustering Iteration 3: Recalculate Cluster Centroids	108
Figure 3.22	<i>k</i> -Means Clustering Iteration 3: Reassign Observations	108
Figure 3.23	Rectangular Versus Hexagonal SOM Grid	109
Figure 3.24	Clustering Countries Using SOMs	111
Figure 3.25	Component Plane for Literacy	112

Figure 3.26	Component Plane for Political Rights	113
Figure 3.27	Must-Link and Cannot-Link Constraints in Semi-Supervised Clustering	113
Figure 3.28	δ -Constraints in Semi-Supervised Clustering	114
Figure 3.29	ε -Constraints in Semi-Supervised Clustering	114
Figure 3.30	Cluster Profiling Using Histograms	115
Figure 3.31	Using Decision Trees for Clustering Interpretation	116
Figure 3.32	One-Class Support Vector Machines	117
Figure 4.1	A Spider Construction in Tax Evasion Fraud	124
Figure 4.2	Regular Versus Fraudulent Bankruptcy	124
Figure 4.3	OLS Regression	126
Figure 4.4	Bounding Function for Logistic Regression	128
Figure 4.5	Linear Decision Boundary of Logistic Regression	130
Figure 4.6	Other Transformations	131
Figure 4.7	Fraud Detection Scorecard	133
Figure 4.8	Calculating the p -Value with a Student's t -Distribution	135
Figure 4.9	Variable Subsets for Four Variables $V_1, V_2, V_3,$ and V_4	135
Figure 4.10	Example Decision Tree	137
Figure 4.11	Example Data Sets for Calculating Impurity	138
Figure 4.12	Entropy Versus Gini	139
Figure 4.13	Calculating the Entropy for Age Split	139
Figure 4.14	Using a Validation Set to Stop Growing a Decision Tree	140
Figure 4.15	Decision Boundary of a Decision Tree	142
Figure 4.16	Example Regression Tree for Predicting the Fraud Percentage	142
Figure 4.17	Neural Network Representation of Logistic Regression	145

Figure 4.18	A Multilayer Perceptron (MLP) Neural Network	145
Figure 4.19	Local Versus Global Minima	148
Figure 4.20	Using a Validation Set for Stopping Neural Network Training	149
Figure 4.21	Example Hinton Diagram	151
Figure 4.22	Backward Variable Selection	152
Figure 4.23	Decompositional Approach for Neural Network Rule Extraction	153
Figure 4.24	Pedagogical Approach for Rule Extraction	154
Figure 4.25	Two-Stage Models	155
Figure 4.26	Multiple Separating Hyperplanes	157
Figure 4.27	SVM Classifier for the Perfectly Linearly Separable Case	157
Figure 4.28	SVM Classifier in Case of Overlapping Distributions	159
Figure 4.29	The Feature Space Mapping	160
Figure 4.30	SVMs for Regression	162
Figure 4.31	Representing an SVM Classifier as a Neural Network	163
Figure 4.32	One-Versus-One Coding for Multiclass Problems	171
Figure 4.33	One-Versus-All Coding for Multiclass Problems	172
Figure 4.34	Training Versus Test Sample Set Up for Performance Estimation	173
Figure 4.35	Cross-Validation for Performance Measurement	174
Figure 4.36	Bootstrapping	175
Figure 4.37	Calculating Predictions Using a Cut-Off	176
Figure 4.38	The Receiver Operating Characteristic Curve	178
Figure 4.39	Lift Curve	179
Figure 4.40	Cumulative Accuracy Profile	180
Figure 4.41	Calculating the Accuracy Ratio	181
Figure 4.42	The Kolmogorov-Smirnov Statistic	181

Figure 4.43	A Cumulative Notch Difference Graph	184
Figure 4.44	Scatter Plot: Predicted Fraud Versus Actual Fraud	185
Figure 4.45	CAP Curve for Continuous Targets	187
Figure 4.46	Regression Error Characteristic (REC) Curve	188
Figure 4.47	Varying the Time Window to Deal with Skewed Data Sets	190
Figure 4.48	Oversampling the Fraudsters	191
Figure 4.49	Undersampling the Nonfraudsters	191
Figure 4.50	Synthetic Minority Oversampling Technique (SMOTE)	193
Figure 5.1a	Köningsberg Bridges	210
Figure 5.1b	Schematic Representation of the Köningsberg Bridges	211
Figure 5.2	Identity Theft. The Frequent Contact List of a Person is Suddenly Extended with Other Contacts (Light Gray Nodes). This Might Indicate that a Fraudster (Dark Gray Node) Took Over that Customer's Account and "shares" his/her Contacts	213
Figure 5.3	Network Representation	214
Figure 5.4	Example of a (Un)Directed Graph	215
Figure 5.5	Follower–Followee Relationships in a Twitter Network	215
Figure 5.6	Edge Representation	216
Figure 5.7	Example of a Fraudulent Network	218
Figure 5.8	An Egonet. The Ego is Surrounded by Six Alters, of Whom Two are Legitimate (White Nodes) and Four are Fraudulent (Gray Nodes)	218
Figure 5.9	Toy Example of Credit Card Fraud	220

Figure 5.10	Mathematical Representation of (a) a Sample Network; (b) the Adjacency or Connectivity Matrix; (c) the Weight Matrix; (d) the Adjacency List; and (e) the Weight List	221
Figure 5.11	A Real-Life Example of a Homophilic Network	224
Figure 5.12	A Homophilic Network	225
Figure 5.13	Sample Network	229
Figure 5.14a	Degree Distribution	230
Figure 5.14b	Illustration of the Degree Distribution for a Real-Life Network of Social Security Fraud. The Degree Distribution Follows a Power Law (log-log axes)	230
Figure 5.15	A 4-regular Graph	231
Figure 5.16	Example Social Network for a Relational Neighbor Classifier	233
Figure 5.17	Example Social Network for a Probabilistic Relational Neighbor Classifier	235
Figure 5.18	Example of Social Network Features for a Relational Logistic Regression Classifier	236
Figure 5.19	Example of Featurization with Features Describing Intrinsic Behavior and Behavior of the Neighborhood	237
Figure 5.20	Illustration of Dijkstra's Algorithm	241
Figure 5.21	Illustration of the Number of Connecting Paths Between Two Nodes	242
Figure 5.22	Illustration of Betweenness Between Communities of Nodes	245
Figure 5.23	Pagerank Algorithm	247
Figure 5.24	Illustration of Iterative Process of the PageRank Algorithm	249
Figure 5.25	Sample Network	254
Figure 5.26	Community Detection for Credit Card Fraud	259
Figure 5.27	Iterative Bisection	261

Figure 5.28	Dendrogram of the Clustering of Figure 5.27 by the Girvan-Newman Algorithm. The Modularity Q is Maximized When Splitting the Network into Two Communities ABC – DEFG	262
Figure 5.29	Complete (a) and Partial (b) Communities	264
Figure 5.30	Overlapping Communities	265
Figure 5.31	Unipartite Graph	266
Figure 5.32	Bipartite Graph	267
Figure 5.33	Connectivity Matrix of a Bipartite Graph	268
Figure 5.34	A Multipartite Graph	269
Figure 5.35	Sample Network of Gotcha!	270
Figure 5.36	Exposure Score of the Resources Derived by a Propagation Algorithm. The Results are Based on a Real-life Data Set in Social Security Fraud	273
Figure 5.37	Egonet in Social Security Fraud. A Company Is Associated with its Resources	274
Figure 5.38	ROC Curve of the Gotcha! Model, which Combines both Intrinsic and Relational Features	275
Figure 6.1	The Analytical Model Life Cycle	280
Figure 6.2	Traffic Light Indicator Approach	282
Figure 6.3	SAS Social Network Analysis Dashboard	293
Figure 6.4	SAS Social Network Analysis Claim Detail Investigation	294
Figure 6.5	SAS Social Network Analysis Link Detection	295
Figure 6.6	Distribution of Claim Amounts and Average Claim Value	297
Figure 6.7	Geographical Distribution of Claims	298
Figure 6.8	Zooming into the Geographical Distribution of Claims	299
Figure 6.9	Measuring the Efficiency of the Fraud-Detection Process	300
Figure 6.10	Evaluating the Efficiency of Fraud Investigators	301

Figure 7.1	RACI Matrix	318
Figure 7.2	Anonymizing a Database	321
Figure 7.3	Different SQL Views Defined for a Database	323
Figure 7.4	Aggregate Loss Distribution with Indication of Expected Loss, Value at Risk (VaR) at 99.9 Percent Confidence Level and Unexpected Loss	331
Figure 7.5	Snapshot of a Credit Card Fraud Time Series Data Set and Associated Histogram of the Fraud Amounts	332
Figure 7.6	Aggregate Loss Distribution Resulting from a Monte Carlo Simulation with Poisson Distributed Monthly Fraud Frequency and Associated Pareto Distributed Fraud Loss	334

Foreword

Fraud will always be with us. It is linked both to organized crime and to terrorism, and it inflicts substantial economic damage. The perpetrators of fraud play a dynamic cat and mouse game with those trying to stop them. Preventing a particular kind of fraud does not mean the fraudsters give up, but merely that they change their tactics: they are constantly on the lookout for new avenues for fraud, for new weaknesses in the system. And given that our social and financial systems are forever developing, there are always new opportunities to be exploited.

This book is a clear and comprehensive outline of the current state-of-the-art in fraud-detection and prevention methodology. It describes the data necessary to detect fraud, and then takes the reader from the basics of fraud-detection data analytics, through advanced pattern recognition methodology, to cutting-edge social network analysis and fraud ring detection.

If we cannot stop fraud altogether, an awareness of the contents of this book will at least enable readers to reduce the extent of fraud, and make it harder for criminals to take advantage of the honest. The readers' organizations, be they public or private, will be better protected if they implement the strategies described in this book. In short, this book is a valuable contribution to the well-being of society and of the people within it.

Professor David J. Hand
Imperial College, London

Preface

It is estimated that a typical organization loses about 5 percent of its revenues due to fraud each year. In this book, we will discuss how state-of-the-art descriptive, predictive and social network analytics can be used to fight fraud by learning fraud patterns from historical data.

The focus of this book is not on the mathematics or theory, but on the practical applications. Formulas and equations will only be included when absolutely needed from a practitioner's perspective. It is also not our aim to provide exhaustive coverage of all analytical techniques previously developed but, rather, give coverage of the ones that really provide added value in a practical fraud detection setting.

Being targeted at the business professional in the first place, the book is written in a condensed, focused way. Prerequisite knowledge consists of some basic exposure to descriptive statistics (e.g., mean, standard deviation, correlation, confidence intervals, hypothesis testing), data handling (using for example, Microsoft Excel, SQL, etc.), and data visualization (e.g., bar plots, pie charts, histograms, scatter plots, etc.). Throughout the discussion, many examples of real-life fraud applications will be included in, for example, insurance fraud, tax evasion fraud, and credit card fraud. The authors will also integrate both their research and consulting experience throughout the various chapters. The book is aimed at (senior) data analysts, (aspiring) data scientists, consultants, analytics practitioners, and researchers (e.g., PhD candidates) starting to explore the field.

Chapter 1 sets the stage on fraud detection, prevention, and analytics. It starts by defining fraud and then zooms into fraud detection and prevention. The impact of big data for fraud detection and the fraud analytics process model are reviewed next. The chapter concludes by summarizing the key skills of a fraud data scientist.

Chapter 2 provides extensive discussion on the basic ingredient of any fraud analytical model: data! It introduces various types of

data sources and discusses how to merge and sample them. The next sections discuss the different types of data elements, visual exploration, Benford's law, and descriptive statistics. These are all essential tools to start understanding the characteristics and limitations of the data available. Data preprocessing activities are also extensively covered: handling missing values, detecting and treating outliers, defining red flags, standardizing data, categorizing variables, weights of evidence coding, and variable selection. Principal component analysis is outlined as a technique to reduce the dimensionality of the input data. This is then further illustrated with RIDIT and PRIDIT analysis. The chapter ends by reviewing segmentation and the risks thereof.

Chapter 3 continues by exploring the use of descriptive analytics for fraud detection. The idea here is to look for unusual patterns or outliers in a fraud data set. Both graphical and statistical outlier detection procedures are reviewed first. This is followed by an overview of break-point analysis, peer group analysis, association rules, clustering, and one-class SVMs.

Chapter 4 zooms into predictive analytics for fraud detection. We start from a labeled data set of transactions whereby each transaction has a target of interest that can either be binary (e.g., fraudulent or not) or continuous (e.g., amount of fraud). We then discuss various analytical techniques to build predictive models: linear regression, logistic regression, decision trees, neural networks, support vector machines, ensemble methods, and multiclass classification techniques. A next section reviews how to measure the performance of a predictive analytical model by first deciding on the data set split-up and then on the performance metric. The class imbalance problem is also extensively elaborated. The chapter concludes by giving some performance benchmarks.

Chapter 5 introduces the reader to social network analysis and its use for fraud detection. Stating that the propensity to fraud is often influenced by the social neighborhood, we describe the main components of a network and illustrate how transactional data sources can be transformed in networks. In the next section, we elaborate on featurization, the process on how to extract a set of meaningful features from the network. We distinguish between three main types of features: neighborhood metrics, centrality metrics, and collective

inference algorithms. We then zoom into community mining, where we aim at finding groups of fraudsters closely connected in the network. By introducing multipartite graphs, we address the fact that fraud often depends on a multitude of different factors and that the inclusion of all these factors in a network representation contribute to a better understanding and analysis of the detection problem at hand. The chapter is concluded with a real-life example of social security fraud.

Chapter 6 deals with the postprocessing of fraud analytical models. It starts by giving an overview of the analytical fraud model lifecycle. It then discusses the traffic light indicator approach and decision tables as two popular model representations. This is followed by a set of guidelines to appropriately select the fraud sample to investigate. Fraud alert and case management are covered next. We also illustrate how visual analytics can contribute to the postprocessing activities. We describe how to backtest analytical fraud models by considering data stability, model stability, and model calibration. The chapter concludes by giving some guidelines about model design and documentation.

Chapter 7 provides a broader perspective on fraud analytics. We provide some guidelines for setting up and managing data quality programs. We zoom into privacy and discuss various ways to ensure appropriate access to both internal and external data. We discuss how analytical fraud estimates can be used to calculate both expected and unexpected losses, which can then help to determine provisioning and capital buffers. A discussion of total cost of ownership and return on investment provides an economic perspective on fraud analytics. This is followed by a discussion of in- versus outsourcing of analytical model development. We briefly zoom into some interesting modeling extensions, such as forecasting and text analytics. The potential and danger of the Internet of Things for fraud analytics is also covered. The chapter concludes by giving some recommendations for corporate fraud governance.

