feedzai

A PRIMER TO MACHINE LEARNING FOR FRAUD MANAGEMENT

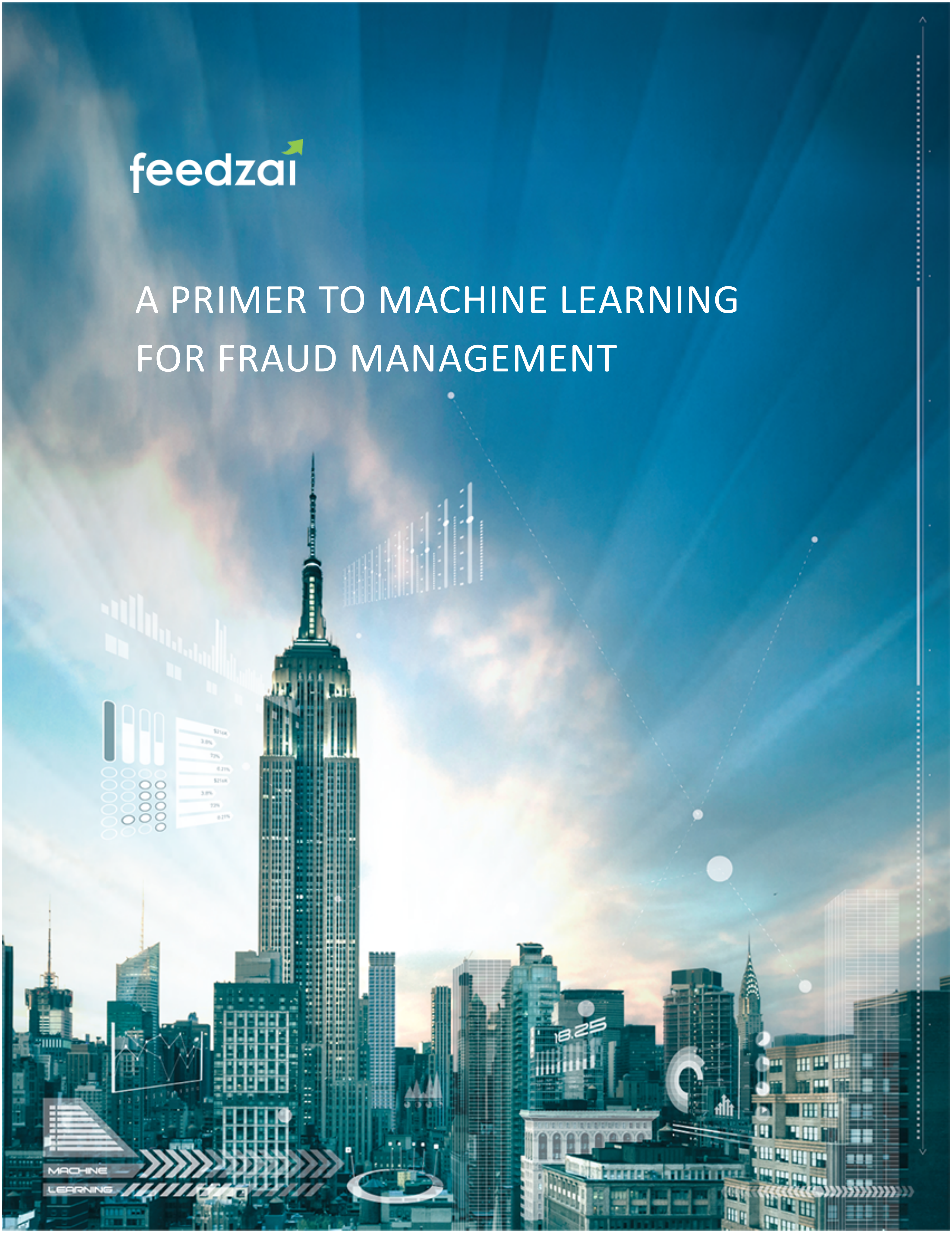


TABLE OF CONTENTS

Growing Need for Real-Time Fraud Identification 3

Machine Learning Today 4

Big Data Makes Algorithms More Accurate 5

Machine Learning For Fraud Prevention 5

Applying Machine Learning 6

Machine Learning Engines 7

Beyond Fraud Prevention 8

Limitations With Machine Learning 8

The Promise of Machine Learning for Fraud Prevention 8

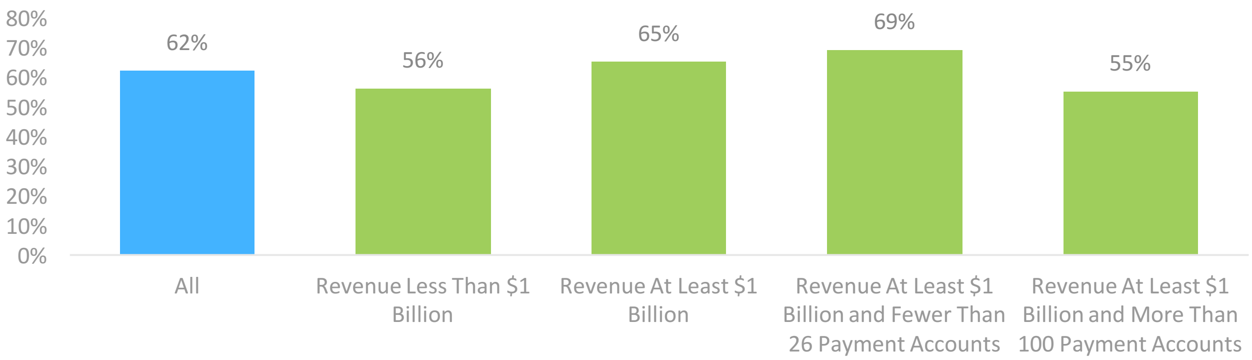
....

GROWING NEED FOR REAL-TIME FRAUD IDENTIFICATION

Fraud attacks are getting to be more sophisticated – as technology evolves fraudsters have elevated their game on payment fraud and money laundering. With access to faster and cheaper computing, fraudsters have shifted their targets to more profitable weaker points in the financial services chain.

Sixty-five percent of organizations with annual revenues of at least $1 billion were victims of payments fraud in 2014 compared to 56 percent of companies reporting annual revenues of less than $1 billion.[1]

Percent of Organizations Subject to Attempted and/or Actual Payments Fraud in 2014



Newer business models are constantly evolving - from instant delivery of goods to virtual cash to digital downloads. However, the growth in opportunities has led to a corresponding growth in online fraud and fraud losses particularly in ecommerce where it is 7 times more difficult to prevent fraud than in the person[2]. According to LexisNexis Fraud Multiplier, in 2015, every $100 of fraud costs a merchant $223 in true costs.

The ever-faster, ever-bigger cycle of attacks leads to a number of consequences:

Magnitudes of attacks are exponentially higher
Fraudsters are employing distributed networks, internal knowledge, big data, and even machine learning to easily detect vulnerability and maximize the size of the attacks

Weakest links create the most exposure
Financial systems are interconnected and consist of a long value chain, a networked ecosystem of multiple entities connecting buyers and sellers. Fraud flows to the least-protected components.

Unexpected attacks can be unsettling and disruptive
Organizations can go from not having a fraud problem to being devastated in just a few days (e.g., Target, Neiman Marcus)

CONCLUSION

Fraud solutions needs to more sophisticated to keep in pace with the fraudsters and react within the short time fraud attacks happen to when they are discovered. Organizations that want to defend themselves against fraud need to have a superior, faster-learning solution that can constantly evolve yet is easy to use and maintain.

MACHINE LEARNING TODAY

Machine learning as a data science to uncover patterns and hidden insights is not entirely a new concept – It has been in play with the use of neural networks starting in the 1980's. The question therefore is, "Why is there a big buzz around machine learning today?"

The answer lies in the fact that advancement in technology and science has enabled game-changing differences in how machine-learning algorithms have evolved and is being applied.

For example, traditionally, human-generated rule sets were the most prevalent approach in fraud management and still continue to be in practice today. But the quantum leap in computing power and availability of big data over the last 5 years has disrupted how data is being used to identify and prevent fraud. Machine learning uses artificially intelligent computer systems to autonomously learn, predict, act and explain without being explicitly programmed. Simply put, machine learning eliminates the use of preprogrammed rule sets - no matter how complex.

Machine learning enables:

Real-time decisions
Advances with in-memory, event streaming technology allow risk scoring and decision making in the sub-second range (i.e., ultra-low latency).

Big Data set processing
Advances in distributed data processing allow analyzing more data while still maintaining real-time decisions without trade-offs between data and latency.

Reduced cycle time
Learning cycles are continuous unlike batch learning where models become out-of-date; With machine learning, the same transactions being scored also update/teach the machine learning models.

Increased effectiveness
Extremely subtle patterns and variations can be detected and delivered (e.g. precision, recall) better than humans in many tasks.

Error-free processing
Enormous amounts of data can now be processed without human-bias or error.

Cost efficiencies
Address long tail "corner case" distribution.

CONCLUSION

Application of machine learning has redefined previous strategies and tools in fraud management delivering benefits that were previously not possible with traditional methods.

BIG DATA MAKES ALGORITHMS MORE ACCURATE

As businesses continues to evolve and migrate to the Internet and as modern money is transacted electronically in an ever-growing cashless banking economy, commerce is increasingly becoming the business of big data science. Of the $11T in US personal consumption expenditures projected in 2017, an astounding 79% of that will be in the form of electronic payments with a face value of $8.5T, or nearly 50% of the GDP of the US[3].

Fortunately, this rapidly expanding "dataverse" also fuels modern artificial intelligence, making big data an inextricable component of today's fraud management. Just like IBM's Deep Blue computer outplayed Garry Kasparov by having learned from millions of chess games, machine learning in general requires access to large amounts of data to be able to learn and generalize knowledge.

Without large amounts of data, a machine-learning algorithm cannot learn. The existence of efficient algorithms to process this data very quickly opened up the possibility for sophisticated machine learning algorithms such as spam detection, efficient content recommendations, autonomous driving cars, image recognition, natural language processing, automatic translation, and of course, fraud management.

MACHINE LEARNING FOR FRAUD PREVENTION

To understand why machine learning is important in fraud management, we need to understand the characteristics of fraud along with the associated business and technical challenges.

Fraud's Unique Characteristics:

Fraud has a long tail distribution
Too many unique cases to pursue.

Fraud patterns change quickly
Slow-learning countermeasures
cannot keep up .

Fraud is adversarial
Professional opponents actively working
to subvert the system at the weakest points .

Fraud mimics good customer behaviors
Good customers are penalized by
over-intrusive countermeasures.

Machine Learning directly addresses many business challenges that are time consuming and expensive – For Example: manual reviews and false positives alone account for almost 40% of the total cost of fraud prevention. According to LexusNexus "The Total Cost of Fraud Prevention" study, merchants allocate as much as one-fourth of costs dedicated to fraud prevention to manual review. Furthermore, new customer channels (e.g., mobile, social), new products and business lines present new risk vectors - fraud through remote channels is up to 7 times as difficult to prevent as in-person fraud.

Machine Learning can:

Reduce manual review queues through fast iterating machine models

Be channel-agnostic

Easily adapt to new business lines using experiential data

Augment human decision-making with increased precision

Reduce false positives with behavior analysis

CONCLUSION
Sophisticated models can reverse engineer machine logic to present human-readable language to explain model decisions.

APPLYING MACHINE LEARNING

Machine Learning models can be used to very efficiently perform **analytics** and deliver **risk scores** in real-time, with greater accuracy by leveraging large amounts of user data. Feedzai's existing model was able to detect +60% of all fraud transactions for a major retailer corresponding to +70% of their fraud money. When trained to include the retailer fraud, the model improved to detect +65% of fraud transactions and +75% of the total fraud money.

Behavior analytics build digital footprints which can then be used to learn from past data in order to make predictions on future, unseen data patterns. For example, in a retail environment, intelligence around user behavior can be used to determine their buying schema – merchandise they buy, stores they frequently visit, times they shop, channel through which they shop etc., Machine learning algorithms can then synthesize this data collected from multiple sources – online and offline - to baseline behavior profiles.   User attributes and other data fields used by machine learning algorithms can automatically learn patterns which are then used to make predictions.

Machine learning can also be used to automatically derive outcome measurements such as a statistical risk (The measurement of the likelihood of incurring loss). The effectiveness of the statistical risk score depends on the model's ability to detect anomalies from known patterns, identify matches to known patterns, and uncover new patterns.

MACHINE LEARNING ENGINES

Mathematical algorithms power machine learning. But, the truth is there is not one single best algorithm that is universally better in all situations - choosing the best algorithm depends on the problem type, size, available resources, etc. Having said that, Random Forests (aka Ensemble of Decision Trees) and Deep Learning have been shown to perform very well in a number of scenarios, with SVM (Support Vector Machines) a close second. Random Forests are more robust for a number of real world problems such as missing data, noise, outliers, and errors. In addition, Random Forests also allow multiple types of data (numbers of different scales, text, Booleans, etc.), can scale very well, parallelize very easily, are fast to train and score, and require less effort to achieve the best results. It is no surprise that Random Forests win many machine learning competitions (as described by Kaggle.com, the world's leading machine learning competition site and data science community).

Algorithm	Pro	Con
Random Forest, aka Ensemble of Decision Trees	• Generalizes patterns well • Robust to different input types (texts, numbers of scales, etc.) • Robust to missing data • Robust to outliers and errors • Fast to train and score • Trivially parallel • Requires less tuning • Probabilistic output (i.e. a score) • Can adjust threshold to tradeoff between precision and recall • Very good predictive power • Found to win many machine learning competitors	• Can become complex to interpret as number of decisions grow (inherent nature of increased capacity to make decisions), but better than all others, especially with Whitebox scoring to demystify decision nodes • Requires labeled data
Deep Learning	• Does not require labeled data • Reduces feature design tasks • Learns multiple levels of representation (e.g. eyes, head, person) • Highly parallel • Very good predictive power, especially in text and image classification problems	• Very slow train, but benefits from recent architecture advances (e.g. GPU's, large clusters) • Cannot handle different input types • Need scaling inputs • Needs tuning • Does not provide probability estimates • Lack of good interpretability • Still missing theoretical foundations
Support Vector Machines (SVM)	• Able to detect non-linear and complex patterns • Effective in very high dimensional spaces • Very good predictive power	• Requires labeled data • Cannot handle different input types • Need scaling inputs • Cannot handle missing values • Not scalable • Slow • Needs tuning • Does not provide probability estimates • Lack of interpretability • Still missing theoretical foundations
Neutral Networks	• Able to represent complex patterns • Good predictive power	• Requires labeled data • Cannot handle different input types • Need scaling inputs • Cannot handle missing values • Not scalable • Slow • Needs tuning • Lack of interpretability
K-Nearest Neighbors	• Robust to missing data • Robust to outliers • Good predictive power	• Requires labeled data • Cannot handle different input types • Need scaling inputs • Cannot handle missing values • Needs tuning • Lack of interpretability

BEYOND FRAUD PREVENTION

Machine learning is not just isolated to identifying and preventing fraud in online retail environment. Machine learning can also be applied wherever large amounts of data can be used to understand and infer behavior for effective decision making.

- Account opening: Validate the authencity of users signing up online to verify and accept more applicants
- Payment authorization: Score payment requests and authorize payments in real-time
- Checkout scoring: Prevent payment chargebacks by scoring transactions during checkout
- Merchant underwriting: Protect your merchant portfolio through merchant underwriting
- Marketplace: Maintain community trust by connecting buyers and sellers

LIMITATIONS WITH MACHINE LEARNING

One of the biggest obstacles to ML is the steep learning curve. Data science knowledge, plus the amount of time and data needed to create models are beyond reach of many risk teams. A steep learning curve means data scientist who do machine learning need to master many different tools such as R, Weka, Python, DBMS, NoSQL data stores, Hadoop jobs, streaming systems and more. Plus, it is very hard to evolve profiles and models to reflect the ever-changing nature of business, e.g. some companies deploy 1-year old models that were trained using 2-year old data.

The second biggest challenge is that a lot of machine learning is grounded on black box decision-making. This is a serious limitation as many policy execution or governance requirements need clear explanations of decisions, e.g. explain to customer why transaction was blocked. Finally, increased capacity to process big data creates an inherent tendency towards include irrelevant data. Machines lack common sense so humans are still needed to supervise.

THE PROMISE OF MACHINE LEARNING FOR FRAUD PREVENTION

While the multiple methodologies in place today to prevent fraud have been successful at keeping fraud rates low for typical payment fraud, the evolving landscape of ecommerce and mcommerce pose newer challenges. These challenges necessitate more innovative solutions that can respond and react quickly to fraud. The need for computational power to process large amounts of data and make decisions real time is imperative for businesses to reach quickly to fraud attacks. Machine learning in this aspect is a promising science that has potential across multiple environments. From payment fraud to abuse, machine learning can easily scale to meet the demands of big data with greater flexibility than traditional methods.

Source:

[1]2015 AFP Payments Fraud and Control Survey

[2] Lexis Nexus True cost of fraud study 2015

[3]http://www.nilsonreport.com/publication_chart_and_graphs_archive.php?1=1&year=2013, "Personal Consumption Expenditures in the U.S.";

US GDP in 2012: $16.2T, http://data.worldbank.org/data-catalog/GDP-ranking-table).

LexisNexus "True Cost of Fraud" 2015 study