EKATA

INTRODUCTION TO MACHINE LEARNING

FOR FRAUD PREVENTION

eBook

TABLE OF CONTENTS

INTRODUCTION

Online fraud is not what it used to be. Not that long ago the most prominent types of online fraud were identity theft and credit card fraud. Identity theft and credit card fraud are still two of the most common types of online fraud. But today online fraud is sophisticated, is happening at lightning speed, and there are many more types. Companies conducting business online are aware of the risk of fraud and chargebacks. But many are not aware of just how fast and sophisticated fraudsters are these days.

TYPES OF ONLINE FRAUD

Fraudsters are constantly changing tactics and finding new ways to commit online fraud. And businesses are facing far more types of online fraud today than they did a decade or so ago. This section lists common types of online fraud.

Identity fraud

Identity fraud is the use of another person's data to deceive or defraud that person or a third party, usually for economic gain. A recent report from Javelin Strategy & Research states that in 2017, 16.7 million U.S. consumers were victims of identity fraud, and fraudsters stole $16.8 billion from U.S. consumers.

Credit card fraud

Credit card fraud is the theft and illegal use of a person's physical credit card or credit card number. Credit card details are obtained by the fraudster without the knowledge of the cardholder. According to the FTC, more than 3.1 million consumer complaints were reported to the Consumer Sentinel Network in 2016. More than 32% of those complaints were about credit card fraud.

Card-not-present (CNP) fraud

Card-not-present (CNP) fraud is when a purchase or payment is made, and the merchant is unable to examine the card used for the transaction visually. According to a recent Javelin Strategy & Research study, CNP fraud rose 40% in 2016 compared to 2015. And a 2017 research report from Juniper Research says that from 2017-2022 retailers are expected to lose $71 billion globally due to CNP transactions.

Middle-of-the-road fraud (card testing)

Middle-of-the-road fraud, also known as card testing, is a relatively new tactic in which fraudsters make purchases in small increments with stolen credit card numbers before moving on to purchasing high-ticket items. By making smaller, less expensive purchases, fraudsters blend in with other consumers making card testing a very effective tactic for fraudsters and a challenge for merchants. According to Radial, credit card testing increased 200% by the end of April 2017, compared to the same quarter in 2016.

New account fraud

New account fraud is where a fraudster opens a new account using the identifying information of another person, typically a person with good credit standing. The accounts are used to purchase products and services or obtain forms of credit like credit cards and gift cards. According to a recent Javelin Strategy & Research report, intermediary new account fraud more than tripled in one year (2016 to 2017), with an estimated 1.5 million victims in 2017.

Account takeover fraud

Account takeover (ATO) fraud is where a fraudster gains access to a consumer's account, usually through stolen login information. Many accounts store customer payment credentials which makes it easy for fraudsters to make purchases. The 2017 Identity Fraud Study by Javelin Strategy & Research found that in 2016, ATO losses reached a total of $2.3 billion, an increase of 61% from the year before.

E-gift card fraud

E-gift card fraud is one of the fastest growing forms of online fraud, and it involves fraudsters stealing the balance of electronic gift card accounts. Fraudsters are using sophisticated techniques to get those gift balances too. Many fraudsters are using botnets to test millions of account number combinations along with stolen pin numbers and passwords. One example of a botnet is "GiftGhostBot" which is capable of testing approximately 1.7 million gift card account numbers per hour, according to a Distil Networks report.

Promo abuse

Promo abuse is where someone gains access to promotions using fake information (email, phone number, etc.) Often fraudsters will use programs that auto-generate phone numbers so that they can take advantage of promotions over and over again. For example, Uber and Lyft often run promotions offering new users a free ride on a car share. Fraudsters will often use many auto-generated phone numbers so that they can receive thousands of dollars in free rides.

Cross-border eCommerce fraud

Cross-border eCommerce fraud is fraud that occurs on transactions involving shipments to foreign locations or fraud that is committed by fraudsters located outside a merchant's country of origin. According to the Radial Annual Fraud Index Report, credit card BIN/IIN country and IP country are red flags for fraud. Internationally, specific geographies represent a higher risk across market segments and verticals such as home, entertainment, cosmetics, and apparel. For example, the IP attack rate within Venezuela is 73.8% for all eCommerce volume for cosmetics.

Other types of online fraud

This section covers only some of the types of online fraud happening today – there's still click/ad fraud, domain spoofing, mobile commerce (mCommerce) fraud, and the list goes on.

CHALLENGES WHEN IT COMES TO FIGHTING ONLINE FRAUD

Online fraud is global and rampant. After all, fraudsters don't limit their efforts to their region only; they commit fraud wherever they can. And thanks to the Internet, fraudsters can engage in fraudulent activities from anywhere, making them hard to catch and be held accountable for their actions. Preventing fraud is challenging for many reasons, one of which is that fraud risk largely depends on the region. When it comes to cross-border eCommerce, some geographical areas are fraught with fraud risk and other areas are less risky. The risk level for a region depends on a number of factors such as payment security methods, population density, and if consumer identity data is available.

Another challenge when it comes to preventing fraud is that consumer identity is regional. Verifying the identity of an individual can be difficult depending on where the person lives. Identity data such as postal codes and phone numbers are not formatted the same way in every region of the world. This presents a challenge for companies that want to verify identities for individuals located in other countries or regions. We recommend a multi-layered approach to verifying identities which means correlating multiple customer identity data points such as email address, phone number, home address, and IP address.

One the most significant challenges for companies conducting business online is that fraudsters are constantly changing tactics and using technologies that allow them to commit many types of fraud quickly. What works to prevent fraud today, may not work tomorrow. Companies need a fraud prevention system that can keep up with fraudsters – and that means a system that uses not only machine learning but also the right data to detect and prevent fraud.

YOU NEED THE RIGHT DATA TO FIGHT FRAUD EFFECTIVELY

Machine learning is an effective tool for fighting online fraud, but machine learning is useless without the right data. Even if you have the latest cutting-edge machine learning (ML) algorithms, you will fail at fighting fraud if you don't feed your ML models the correct data.

Helping businesses improve their machine learning fraud models is one of the reasons we invest and focus heavily on the best global identity data and network, sophisticated data science, and enterprise cloud infrastructure.

Ekata Identity Network

The Ekata Identity Network allows us to analyze millions of historical transactions across our customer base and use this intelligence for any new transaction.

The intuition behind the Ekata Identity Network is simple – there are predictable patterns that correspond to the behavior of genuine customers and fraudsters. Since we are used in their workflows by hundreds of customers processing millions of transactions every single day, we have access to this unique repository of information that is extremely valuable.

The Ekata Identity Network includes multiple identity element velocities, transactional frequencies, and linkage history attributes. Principally, each attribute represents a different behavioral pattern that we believe has value in identifying the nature of a transaction.

Ekata Identity Graph

Our Identity Graph is a fully-integrated, high-availability database of identity data that has been curated and corroborated to deliver unparalleled coverage, accuracy, and performance. The real power is the linkages that connect people, phones, addresses, emails, and IPs to help businesses confidently assess and verify consumer identities.

Individual data attributes are important, but linkages are exponentially more powerful, especially when it comes to verifying a new customer where a company has no history or data. Our Identity Graph houses over 5 billion global contact records and over 8 billion linkages to connect these consumer data attributes.

Ekata Identity Check API

Identity Check API helps businesses get a clearer picture of their customers by leveraging real-time global data, machine learning, and network insights across the five core consumer data attributes of email, phone, person, address, and IP. For example, Identity Check could verify that the email matches the name, or tell you information about the email – e.g. email first seen date, disposable status, etc. Identity Check returns 70+ data signals, leverages proprietary Identity Network insights, and provides a Confidence Score in 250 milliseconds.

Ekata Confidence Score

The Confidence Score provides a comprehensive assessment of each transaction by leveraging the millions of patterns across our Identity Network and Identity Check's 70+ data signals. This assessment is delivered as a single score on a 0-500 range. The Confidence Score can be accessed in three ways: from a fraud platform that has Identity Check directly integrated (see platform integrations), as an attribute in the Identity Check API, and in the Ekata Insight manual review solution.

Ekata data is designed for risk

Machine learning is how a fraud detection system keeps up with fraudsters but feeding ML models the right data is how a fraud detection system effectively catches real fraud.

Ekata data is explicitly designed for risk. We provide a suite of identity data APIs that are designed to improve the performance of fraud detection models. And our data is comprehensive, global, consistent, and reliable – features are not changed in a way that would break your ML models.

MACHINE LEARNING 101

Once you have the right data, you can start leveraging machine learning to fight online fraud.

This section covers the basics of machine learning including some common terms and concepts. But first, why is machine learning so popular these days?

What is AI?

Artificial intelligence (AI) is an area of computer science that involves giving machines the ability to perform tasks that typically require human intelligence, tasks such as speech recognition, visual perception, and reasoning.

What is machine learning?

Machine Learning is a subset of AI and involves giving computers the ability to learn without being explicitly programmed. Andrew Ng, Landing.ai founder and CEO, and adjunct professor at Stanford University, describes machine learning as "technology that lets a computer get smarter and smarter all by itself just by looking at data."

Why machine learning is popular now

Machine learning has been around for decades, but it is only in the past six years or so that the use of machine learning has taken off. Machine learning is just about everywhere now – on smartphones automatically tagging photos and powering the voice-activated intelligent assistant. On eCommerce websites and mobile apps recommending products and preventing fraudulent transactions. In connected cars monitoring vehicle performance and powering methods for predictive maintenance.

Machine learning is popular today for many reasons, but the main reasons are the availability of computational power, access to high-quality training datasets, and an abundance of data.

Today, major tech companies like Amazon, Google, and Microsoft provide computing services that allow companies to build, train, test, and deploy highly scalable ML models. Any company can take advantage of machine learning. Access to computational power is no longer limited to big government agencies and educational institutions.

Another reason for the rising use of machine learning is the vast volumes of data available today. Thanks to the internet of things, data is exploding – billions of connected devices are generating massive volumes of data every second. And this data is being generated by both humans and machines (M2M). Companies are using machine learning to collect, analyze, and leverage this abundance of data.

Now that we've highlighted why machine learning is popular today, the rest of this section covers a number of common ML terms and concepts.

Model vs. algorithm – what's the difference?

Machine learning models and algorithms are closely related, but they are two different things.

Algorithm

Algorithms are the "learning" part of machine learning. In most types of machine learning, an algorithm is fed training data in order to learn patterns and values. Choosing an algorithm depends on a number of factors such as:

- Computational resources and time available
- Size and quality of the data
- How fast the task must be completed
- Level of accuracy required
- If explainability is needed
- If the algorithm makes use of linearity

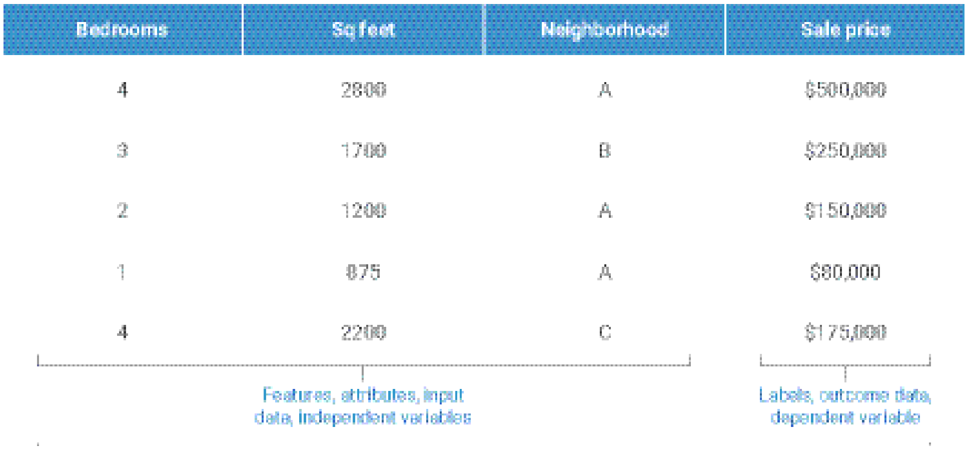
The Microsoft Azure documentation includes a section about "How to choose algorithms for Microsoft Azure Machine Learning." The documentation is comprehensive, and there is a nice "algorithm cheat sheet" infographic.

Model

A machine learning model is the part that comes after the learning process of an algorithm. Once an algorithm has been trained, a model is saved on what the algorithm has learned. The model is then used later to make predictions on new data.

Companies build an ML model to address a specific business problem, predicting the value of a house for example.

Training data for a house price algorithm

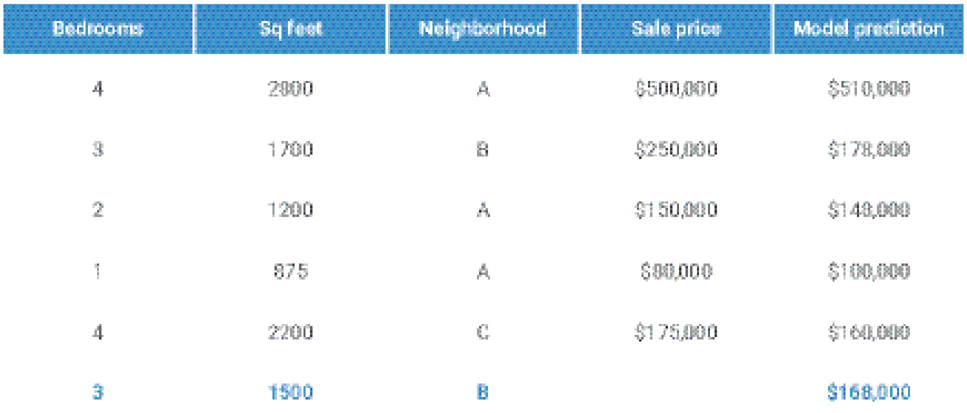


Bedrooms	Sq feet	Neighborhood	Sale price
4	2800	A	$500,000
3	1700	B	$250,000
2	1200	A	$150,000
1	875	A	$80,000
4	2200	C	$175,000

Features, attributes, input data, independent variables
Labels, outcome data, dependent variable

Predict the value of a house

If we use a machine learning algorithm, we could come up with something like Sales price = number of bedrooms X 0.84 + square feet X 0.12 + neighborhood X 2.32 + 201.23



Bedrooms	Sq feet	Neighborhood	Sale price	Model prediction
4	2800	A	$500,000	$510,000
3	1700	B	$250,000	$178,000
2	1200	A	$150,000	$148,000
1	875	A	$80,000	$100,000
4	2200	C	$175,000	$160,000
3	1500	B		$168,000

How does the computer figure out the above relationship?

You will typically see two families of models in the fraud prevention space: regression and tree-based. The above house price model is an example of a regression model. ML model types are explained later in this next section.

ML training methods come in many forms such as supervised learning, unsupervised learning, deep learning, and reinforcement learning. The most commonly used type of machine learning training method is supervised learning.

Supervised machine learning

In supervised learning, algorithms are provided labeled training examples from which the algorithms learn patterns and values. The training data includes known inputs and outputs. Each training example is labeled with a value of interest, a "dog" or "cat" for instance. The algorithm looks for patterns in the value labels. Once the algorithm finds the best pattern, it uses that pattern to make predictions. For example, an algorithm would be trained to detect if a cat is in a photo. In supervised learning, the algorithms are taught the right answers from known data.

The goal of supervised learning is for machines to map a function based on the training examples given to them and then make predictions without requiring explicit instructions. Most of the companies in the fraud space, including Ekata, are using supervised learning.

What machine learning can do

Most businesses today that are gaining practical value from machine learning are using supervised learning. The below chart is a simplification of the supervised learning concept and shows some of the applications of supervised learning:

Input A	Response B	Application
Picture	Is there a cat? (0 or 1)	Photo tagging
Picture	Is there offensive content? (0 or 1)	Content moderation
English sentence	French sentence	Language translation
Audio clip	Transcript of audio	Speech recognition
Vehicle camera or sensors	Position of other objects	Autonomous vehicles
Hardware or machine sensors	Is it about to fail?	Preventive maintenance
Loan application	Does applicant have good credit?	Loan approvals
Email	Is this spam? (0 or 1)	Email spam protection
Phone number	Is this a valid number? (0 or 1)	Phone number validation
Text	Positive, negative, or neutral?	Sentiment analysis

Types of supervised learning models

Classification

Classification is about predicting a label. For example, predicting that a given image contains a cat.

Regression

Regression is about predicting values, like our house price model example earlier in this section.

Logistic regression

Logistic regression is used to solve problems with two class values (binary classification).

Linear regression

Linear regression means that the model expects that the input variable(s) and the output variable have a linear relationship.

Linear regression takes the format:

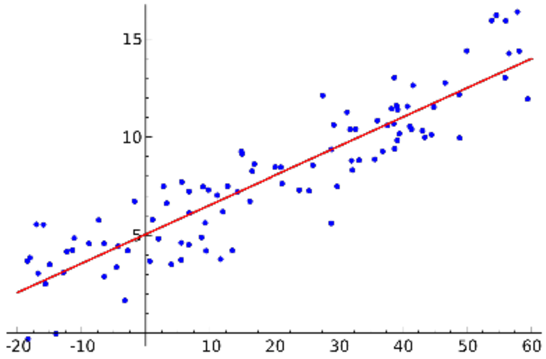
Answer =

(Weight1 * Feature1)

+ (Weight2 * Feature2)

+ …

+ (WeightN * FeatureN)



E.g.

avg_home_price =

2.5*median_annual_income

+ 102.0*square_footage

- 100,003.7*miles_from_downtown

Tree-based models

Tree-based algorithms are usually the better choice for fraud prevention. We explain why tree-based models are a popular choice for fraud prevention in a separate blog post.

Decision tree

Decision trees consist of a series of checks that each branch in different directions. Trees can branch based on:

- Numeric thresholds
- Whether a field has some value
- Where any other condition is met

Decision forest

A decision forest is an ensemble model that is primarily used for classification but can also be used for regression. This type of ML model builds multiple decision trees, learns from labeled data during the building process, and then votes on the most popular output class. Voting is one of several methods for combining predictions from different models.

Ensemble models typically provide higher accuracy and broader coverage than single decision trees because ensemble models involve building multiple models instead of a single model. The models are combined into one generalized model that is less impacted by training data outliers.

Random forest (or random decision forest)

Random forest is an ensemble model that builds multiple decision trees and uses bagging to combine the predictions of each tree (the model averages the predictions of each tree).

Unsupervised learning

In unsupervised learning, algorithms are given input data but no output data. Algorithms learn to identify patterns on their own; they are not trained with labeled training data like supervised learning algorithms. Anomaly detection and clustering are examples of unsupervised learning techniques.

A well-known example of unsupervised learning is Google training a neural network to identify cats. Google's system learned how to identify cats from viewing unlabeled YouTube videos. This learning system was built back in 2012 and was comprised of a network of 1,000 machines. At the time, the network had more than 1 billion parameters that were trained on 16,000 CPU cores. Unsupervised learning requires a lot of computational power.

Deep learning

Deep learning typically refers to a class of algorithms that are a method of learning in neural networks. A neural network is a computer system that is loosely based on how the human brain works. Deep learning algorithms learn feature representations automatically which means less, if any, time must be spent on feature engineering.

Deep learning is an area of machine learning, and deep learning algorithms are often used to enhance computer vision techniques such as object recognition, motion analysis, and scene reconstruction. Deep learning can also be applied to other input types like audio. Many different types of neural networks exist today, and convolutional neural networks (CNN or ConvNet) are the type of neural networks typically associated with image processing tasks like the ones mentioned above.

Note that deep learning algorithms can be trained in a supervised, unsupervised, or semi-supervised manner. Supervised is the most common training method, however.

Reinforcement learning

The goal of reinforcement learning is to refine the ML system so that it is driven to choose the right priorities in specific situations. This type of machine learning typically involves an environment that is formulated as a Markov decision process (MDP). MDP is a mathematical representation of a modeling decision-making process in which outcomes are partly controlled by both a decision maker and at random. This type of machine learning is commonly used in the field of robotics which includes autonomous vehicles.

Quantifying model performance

Once you've built an ML model, you need to determine how well the model is performing. Assessing the performance of unsupervised learning algorithms is far more difficult than supervised learning algorithms. Because unsupervised learning does not use labeled training data, it's difficult to know if the results are accurate. However, if a supervised learning algorithm is trained to identify a cat in a photo, and it instead identifies a dog, then you know it's not performing as expected.

In supervised learning, the accuracy and performance of models can be measured and refined. Among the most commonly used methods to evaluate model performance are Receiver Operating Characteristic (ROC) and Kolmogorov-Smirnov (KS). Most of our customers use the ROC method, so this is the method we will cover here.

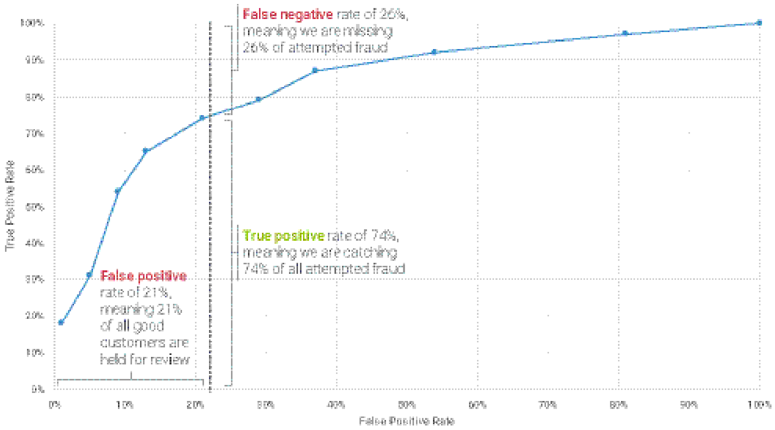
Receiver Operating Characteristic (ROC)

ROC is a plot graph that compares the true positive rate against the false positive rate at different thresholds. For example, if you want to evaluate the performance of a fraud detection model, you could assess transactions that are flagged as fraud and held for manual review.

	Model flagged	Model didn't flag
Actually fraud	**True positive**	**False negative**
Actually good	**False positive**	**True negative**

In this scenario, a true positive would mean that the transaction was flagged correctly and sent to manual review – that the transaction is indeed fraudulent. A true negative would mean that the transaction was not flagged and that this is indeed a good customer. A false positive would mean that a transaction for a good customer was flagged incorrectly as fraud and held for manual review. Finally, a false negative would mean that a truly fraudulent transaction was not flagged by the system as fraud – a fraudulent transaction was allowed to go through the system.

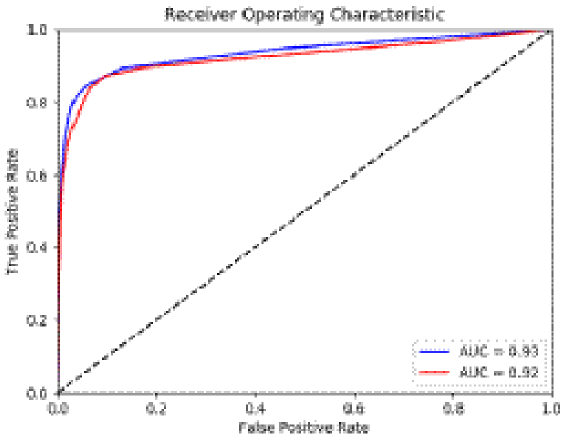
The ROC method helps you understand the actual business impact of what your model is doing and how much true fraud you're catching. The goal of the model in the above scenario is to have the least percent of good customers impacted by the fraud detection system and a higher percentage of true fraud flagged. A perfect result would be if 0% of good customers are impacted and the system catches 100% of true fraud. On the chart, the dot would be located at the very top corner – 0.0 and 100%.



Area Under Curve (AUC)

Another term you'll often hear regarding model performance evaluations is Area Under Curve (AUC). The AUC tells you how well your model is performing overall, but not much about how well the model is solving your specific business problem. The curve is just an example of how the model is doing in general, not how the entire business is doing.

The AUC can tell you if changes made to your model improves or worsens the results of the model. ROC curves drive the number of reviews and the number of chargebacks. If you know the percentage of chargebacks and the percentage of true fraud captured, you can come up with a dollar value of the money saved because of fewer chargebacks.



MACHINE LEARNING CAN KEEP UP WITH FRAUDSTERS

If your company is conducting business online, then you likely have fraud prevention strategies in place to keep the fraud out, while moving more good customers through the approval process. But what fraud prevention system do you have, and can it keep up with the sophistication and scale at which fraud attacks are taking place? A recent IBM Security Intelligence blog post reports that in the first quarter of 2018 alone, there were 210 million attempted fraud attacks.

Are you using a rules-based system, or one driven by machine learning, or maybe a combination of both? Any fraud prevention system that relies heavily or solely on human input will not be as effective at fighting fraudsters who are using vast networks and automation to commit fraud. Machine learning models can provide scale as the ML algorithms can be trained and refined to see patterns in data and transactions that humans cannot.

The bottom line is that machines are more efficient and better suited to perform many types of repetitive tasks than humans. And preventing fraud involves numerous repetitive tasks that must be completed faster than humans can manage.

MACHINE LEARNING POSITIVELY IMPACTS BUSINESS OUTCOMES

With machine learning, companies can catch more fraud, improve customer experiences, and change the lives of fraud prevention teams for the better. Machine learning allows businesses to automate the fraud prevention process and optimize workflows. You can also better ensure that your fraud detection system captures a higher percentage of genuinely fraudulent transactions and that the number of false positives is significantly reduced.

Need help adding machine learning to your rules-based system or quality identity data to your fraud model to improve your AUC? Our team of machine learning solutions architects can help. Contact us today, or request access to our Ekata Machine Learning Guides for Identity Check API and Reverse Phone API.