WHITE PAPER

More Data More Insight

What does more data mean for boosting your fraud prevention?

Samira Golsefid, Ph.D.

PayPal

Contents

Overview

At PayPal, we often talk about the need for machine learning-powered systems to combat a new wave of sophisticated digital fraud. But what are the requirements for using machine learning? In this paper, we will briefly review common fraud use cases and typical data fields that we need for each of these use cases. We will also discuss the impact that size and volume of data has in creating a powerful model that helps us detect fraud patterns and we will summarize the minimum data requirements to start machine learning modeling.

In the first section of this paper, we will cover the general machine learning flow. In the next section, we will explain what kind of input data is needed to build a machine learning model and will present the type of data and resources for solving different fraud uses cases. In the third section, we will describe several factors that impact the performance of the machine learning process, such as the size of the data set, the number of bad behaviors, the time frame, and the rate of data flow.

Machine Learning Flow

Machine learning is an algorithm or model that learns patterns in a historical data set and then predicts similar patterns in a new data set. At PayPal, we use both rules and machine learning models to help predict fraudulent behaviors. Rules help us detect scenarios that frequently occur, such as “decline transactions in a specific region.” They can also be extracted from historical data, for example as “decline a transaction if the number of bad transactions per device related to this transaction is more than 10 in the past month.”

When the number of scenarios increases significantly, and the patterns become more complex, machine learning offers a more practical approach. It enables us to create algorithms to process large datasets with many variables and help find correlations with different features. Machine learning is faster, can cope with structured and unstructured data, and is easy to retrain and update with the latest data. There’s also less manual work involved, helping reduce operational costs.

Machine learning is essentially a subset of artificial intelligence that enables us to learn without explicitly defined conditions with predefined rules. The machine learning models need to be taught what the pattern of bad behavior is to be able to expose that to the new data. For example, for detecting fraudulent transactions, it would mean that instead of setting up precise rules for what constitutes a fraudulent transaction, data scientists can feed a machine learning algorithm thousands of examples of good and bad transactions. The machine learning model then finds patterns in these transactions and uses those patterns to identify future bad behaviors.

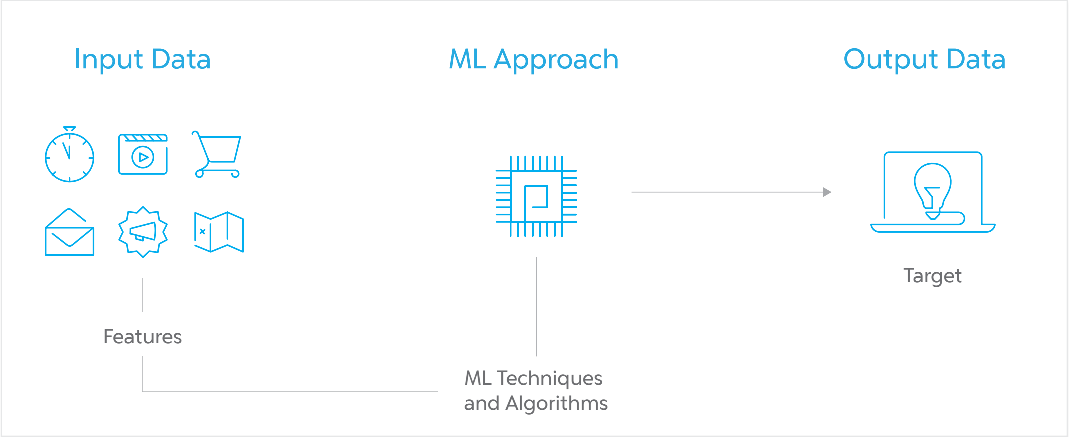


Figure 1: General Machine Learning Flow

A machine learning system has three main elements:

- **Input Data**: A wide variety of data can be used as an input for machine learning purposes. For fraud-related use cases, we may use a wide variety of data, including device, IP, email, phone, address, and transaction data.
- **Learning Process**: There are two main approaches in training a machine learning model: supervised and unsupervised learning. In supervised learning, observations contain input/output pairs. These sample pairs are used to "train" a machine learning system to recognize specific patterns for correlating inputs to outputs. In unsupervised learning, those labels are omitted. In this form of machine learning, rather than being "trained" with sample data, the machine learning system

finds structures and patterns in the data on its own. Supervised machine learning models are used more commonly to respond to a whole range of fraud domain use cases.

- **Output Data**: Machine learning can be used to deliver results that are either predictive or prescriptive (suggesting an action) — provided as a report or integrated response into other enterprise systems. For instance, to decline or approve transactions as a result of running the machine learning model.

In the next sections, we will describe what kind of data is required when building a machine learning model for different fraud use cases.

What Kind Of Data Do We Need?

Fraud is continuously evolving and threatens companies in different ways. The main question here is, “what kind of data do we need to build a machine learning model for different fraud use cases?” In this section, we will review common types of fraud and what kind of data we need.

New Account Origination

New Account Origination (NAO) is one of the fastest-growing types of fraud. NAO fraud occurs when a fraudster uses stolen credentials or creates fictitious identities to open a new bank or credit card account, submit an application, or request a quote for a loan, insurance, or other virtual products. This type of fraud can result in stolen funds, products, and lost revenue. Fake accounts can be opened either by using stolen credentials (third-party fraud[1]) or by creating fake identities (a category of first-party fraud).

It is critical to stop fraudsters at account creation to help minimize losses rather than spending more resources later on costly chargebacks. NAO fraud is hard to catch; unlike authenticating a returning user, account creation is the first time a company

[1] **First, Second, and Third-Party Fraud**
First-party fraud: Mostly seen in financial-related fraud, first-party fraud occurs when borrowers trick a lender into believing they have good credit to borrow money they do not intend to pay back. They may do this by using their own data or building a fake persona and then slowly building this fake persona’s credit over time.
Second-party fraud: Often considered first-party fraud, Second-party fraud is a type of application fraud in which the fraudster is a friend or trusted acquaintance of the party whose name is on the application. Again, the credited money is not paid back.
Third-party fraud: Often called identity theft or identity fraud, third-party fraud occurs when someone creates a fraudulent account or application in someone else’s name while pretending to be that person. However, unlike second-party fraud, the perpetrator is unknown to the victim and has most likely acquired the victim’s identity illegally.

sees the new client, and there is nothing to compare them to. At the time of account creation/enrollment, the interaction of the fraudsters is minimal with the system. Therefore, it is challenging to detect a fake user at signup in real-time.

PayPal helps enable organizations to build scoring machine learning models to automate this detection process. To develop models and rules that help identify fraudulent accounts, we heavily rely on device data, third-party information, session analysis, and information collected during the enrollment process. We build hundreds of signals on top of the device, email, IP, phone, user, and address information that feeds our model to help detect suspicious behavior. We can create a signal to help detect a mismatch between a user's real and stated location, or abnormal phrases when typing familiar information such as a first and last name. We also leverage third-party data for rate risk factors around the identity attributes for authentication purposes:

- **Device Fingerprint**: Uniquely identify a desktop, tablet, or mobile device used to create a new account based on browser type, device type, browser language, or operating system (OS), among others with Device Recon.
- **Identity Scores**: Get risk scores and reason codes instantly to help prevent fraudulent users from creating accounts using synthetic or fake identities.
- **Financial History**: Incorporate users' financial credibility into the scoring or Early Warning. The more history there is, the better the results.
- **Public Records**: Reverse lookups on addresses, names, and phones.

- **Email Address**: Identify if the email used to sign up for the new account is risky or blacklisted by comparing it with external databases.

Some third-party sources provide data points that can be used for authentication purposes.

Account Takeover

Account Takeover (ATO) occurs when an unauthorized person acquires a legitimate user's sensitive logon data to gain access to an existing account that does not belong to them, and changes account details such as login credentials or personal information. A successful account takeover attack can lead to fraudulent transactions, unauthorized shopping using a victim's compromised account, or illegal transfers to other accounts. With access to a legitimate user's account, the perpetrator can steal confidential information, such as their driver's license or Social Security number, and can pretend to be that person or sell the information to other criminals.

While there are multiple reasons behind why an account is taken over, more often than not it is for monetary gain. The direct cost of an account takeover is evident, from fraudulent transactions and bank transfers to the purchase of goods through a compromised ecommerce account. The are many consequences for businesses affected by ATO, but the most damaging ones are increased chargebacks, increased customer transaction disputes, loss of customer trust, customer churn, and brand damage. Often, without a customer claim, a company may not realize account takeover has occured.

Account Takeover affects multiple account types and various ecommerce transactions. Credit card, checking, savings, and brokerage accounts are at risk for account takeover, which can start days, weeks, or months before a customer becomes aware of fraudulent activity. We need to be proactive instead of reactive to achieve account takeover detection and prevention. An account takeover solution must detect activity or other signals that are abnormal for a user before the takeover happens, such as trying to access an account several times in a short amount of time, or thousands of new accounts created from the same IP, location, or type of device.

PayPal applies machine learning models and rules to assess users' behavior patterns in real-time to detect if they are acting consistently, or if their behavior points to malicious intent. Key data points that help reveal fraud include device, email, IP, phone, transaction, and user information. We create hundreds of signals to capture high-risk activity, such as the number of times an email, phone number, or password associated with an account are changed in a day. We can create a signal to count the number of login attempts, triggered if thousands of username and password combinations are tested each second. We can also create a signal that triggers if the number of login attempts is higher than usual for this user or across an industry. Additionally, we can check if an account holder made similar changes in the past or made changes to the account that matches a pattern of account takeover. Third-party data helps us to check if the new information being added to an account—such as an email or phone—has a history of being high-risk. Third-party data also helps us to uncover multiple high-risk changes:

- **IP Geolocation**: Identify the geographic origin of a login based on the IP details of the user's browser.
- **Device Fingerprint**: Uniquely identify a desktop, tablet, or mobile device based on the browser type, device type, browser language, OS, etc.
- **Public Records**: Reverse lookup of addresses, names, and phones to validate the user logging in.
- **Proxy Detection**: Recognize when hackers are trying to bypass geolocation controls by using proxies to spoof their IP address.

Transaction Fraud

Transaction fraud is one of the most common types of fraud. In the payments industry, transaction fraud occurs when a fraudsters illegally uses a credit card/card number without the real cardholder's knowledge. Once a cardholder sees a transaction they did not make on their credit card statement, they have the right to dispute the charge by contacting their bank. The bank or credit card company conducts an investigation and returns the money to the cardholder.

Payment fraud is a fraudulent or unauthorized transaction completed by cyber-criminals. Fraudsters sometimes initially make several small transactions to make sure a stolen card has not been blocked before using it for a large transaction. In some cases, they will attempt a large initial purchase. The machine learning model can detect these kinds of patterns by comparing them to the genuine user's spending behaviors.

At PayPal, we look at a wide variety of information, such as transaction, device, email, IP, phone, user, and address data to help detect whether a transaction is legitimate. We create signals to help us identify suspicious transactions, like a shipping-billing address mismatch, unusually large orders (e.g. a large number of items or a large number of the same item), using different cards for multiple orders with the same shipping address, ordering rushed or next day shipping, or trying different expiration dates after the initial decline. While these are not necessarily suspicious on their own, they, along with other red flags, help us to detect suspicious behavior. By linking transaction information to additional third-party data, we can help businesses spot anomalous behavior and identify high-risk indicators in real-time:

- **IP Geolocation**: Identify the geographic origin of a login based on the IP details of the user's browser.
- **Issuing Bank Data**: A Bank Identification Number (BIN) identifies the institution and their location to prove the authenticity of the card used to make transactions.
- **Device Fingerprint**: Uniquely identify a desktop, tablet, or mobile device based on browser type, device type, browser language, and OS, among others with PayPal data.
- **Proxy Detection**: Recognize when hackers are trying to bypass geolocation controls by using proxies to spoof their IP address.
- **Email Address**: Identify if the email used for the transaction was recently created, risky, or blacklisted by comparing it with external databases.

Deposit Banking

Fraudsters exploit the channels designed to improve customer experience, gathering details on financial institution operations to defraud a bank.

Mobile Remote Deposit Capture

Mobile Remote Deposit Capture (MRDC) is a kind of check deposit fraud where consumers can deposit a check by simply taking a photo of it with their mobile phone. The ability to deposit checks via mobile devices is one of the most sought-after mobile banking app features. While not as prevalent in the medium-to-large-sized banks due to existing solutions like EWS and MiTek, newer challenger banks are facing increased challenges due to people double-dipping or even triple-dipping their checks as a result of current check clearing practices. Understanding a user's behavior, such as the time they usually deposit, thier location, device, and other characteristics helps detect MRDC fraud.

A platform will be configured to analyze each transactional event specific to MRDC. The event will be analyzed based on current IP, device, and transaction (amount, time, frequency, etc.) data collected and then compared to previous events. The event is then scored using a set of transactional rules, along with one or more machine learning models. The results of this scoring will be used to define auto-decisions, which may include recommendations such as "accept," "reject," and "review" transactions before the money is moved.

Card-Not-Present Fraud

Card-Not-Present (CNP) fraud is another type of fraud that includes telephone, fax, Internet, and mail-order transactions where the cardholder does not physically present the card to a merchant. Most CNP fraud involves the use of card details that have been obtained through skimming, hacking, email phishing campaigns, or telephone solicitations. Since both the card and cardholder aren't physically present, it can be difficult for merchants to verify the purchaser's identity. The two most commonly used methods for authenticating online transactions are card verification numbers (CVN) and negative lists (also known as blocklists). The address verification system (AVS) is also an effective way to verify the address of the person claiming to own the credit card. The system checks the billing address of the credit card provided by the customer with the address on file at the credit card company.

Card-Present Fraud

A transaction is only considered as card-present (CP) if payment details are captured in-person at the time of sale. This type of fraud may involve the use of the actual stolen card or a fraudulent duplicated card made using a genuine card's number and magnetic stripe information. CP fraud is theoretically more straightforward to prevent than CNP fraud because the merchant has the opportunity to examine the credit card and a customer's behavior for signs of suspicious activity.

Automated Clearing House

Automated Clearing House (ACH) transactions are electronic fund transfers between bank accounts using a batch processing system. It is a popular alternative to paper checks and credit card payments. ACH fraud occurs when unauthorized funds are transferred into a bank account. This fraud can occur if someone gets access to an account holder's account number and bank routing number. The funds can then be used for fraudulent activity, potentially leading to a significant loss for both consumers and businesses.

Wire Transfer

Wire transfer is an electronic transfer of money like ACH payments, but it is generally faster. Wire fraud can take place in the form of account take over or first-party fraud. The quick payment clearing time, which is much faster than ACH or checks, makes this method particularly attractive to fraudsters. Unlike ACH, funds cannot be reversed in case of wire fraud.

How Much Data Do We Need?

“How much data do we need?” is probably the most common question asked when building a machine learning model. Unfortunately, there is no easy answer to this question. Our usual answer is “as much as possible” because the more data we have, the better we can identify the structure and patterns that are used for forecasting. The quality and amount of training data is often the single most dominant factor that determines the performance of a model. How much training data do we need to train a model? The answer is that it depends on different factors, including the use case we are trying to solve for, the level of desired performance we want to achieve, the input features we have, the noise in the training data, the noise in our extracted features, and the complexity of our model, among others. We describe the most important factors that affect the quantity of data needed for the model:

Labeled Data

Supervised machine learning models are successfully used to respond to a whole range of fraud use cases. However, these models are data-hungry and their performance relies heavily on the size of training data available. Supervised Learning is based on the availability of high-quality labeled data. We need to mark fraudulent behavior in historical data to be able to train our model to identify similar fraudulent patterns in the future. To put it more technically, labeling our training data gives a model the ability to correctly predict, classify, and otherwise analyze data to help generate a meaningful output. It is hard to determine how many “bad” labels we need to have a high-performance machine learning model, but the rule of thumb is that we need at minimum between 1000 and 5000 examples of bad transactions. Again, the more data we have, the more advanced the machine learning model we can develop to help identify complicated fraudulent patterns.

The Number Of Input Features

One of the most important factors is the number of input features. The more complex the problem, the more data we need. Generally, the more dimensions data has, the more data we need. By increasing the number of features, the complexity of patterns in the data set increases, and we need more data to detect these patterns. In the fraud domain, we use high dimensional data sets to train a model. At the transaction event, we may have around a hundred raw data points and we can build thousands of features on top that help describe a user’s behavior. Therefore, there is an exponential increase in the difficulty of the problem as the number of input features is increased and we need more data to find the patterns. It is always necessary to have more records than features. As data exhibits a lot of random variation in the most practical of applications, it is usually essential to have many more records than features. From a purely statistical point of view, it is always required to have more observations than parameters.

The Complexity Of The Machine Learning Algorithm

Another factor is the complexity of the learning algorithm that you are going to use for a given problem. The amount of data you need also depends on the complexity of your chosen algorithm. Typically, detecting a fraud pattern cannot be easily separated by a single line between fraudulent behavior and good behavior. We need to select a more robust classifier, such as random forest to identify the complex relationships. Generally, a model with higher dimensional data (high number of features) requires more training data to train appropriately. Increasing the volume of data can help improve our machine learning model and keep it up to date while training it for our business purposes. Algorithms can continue to improve in skill as we give them more data. It depends on the problem: the more complex it is, the more data you need.

Comprehensive Data Set

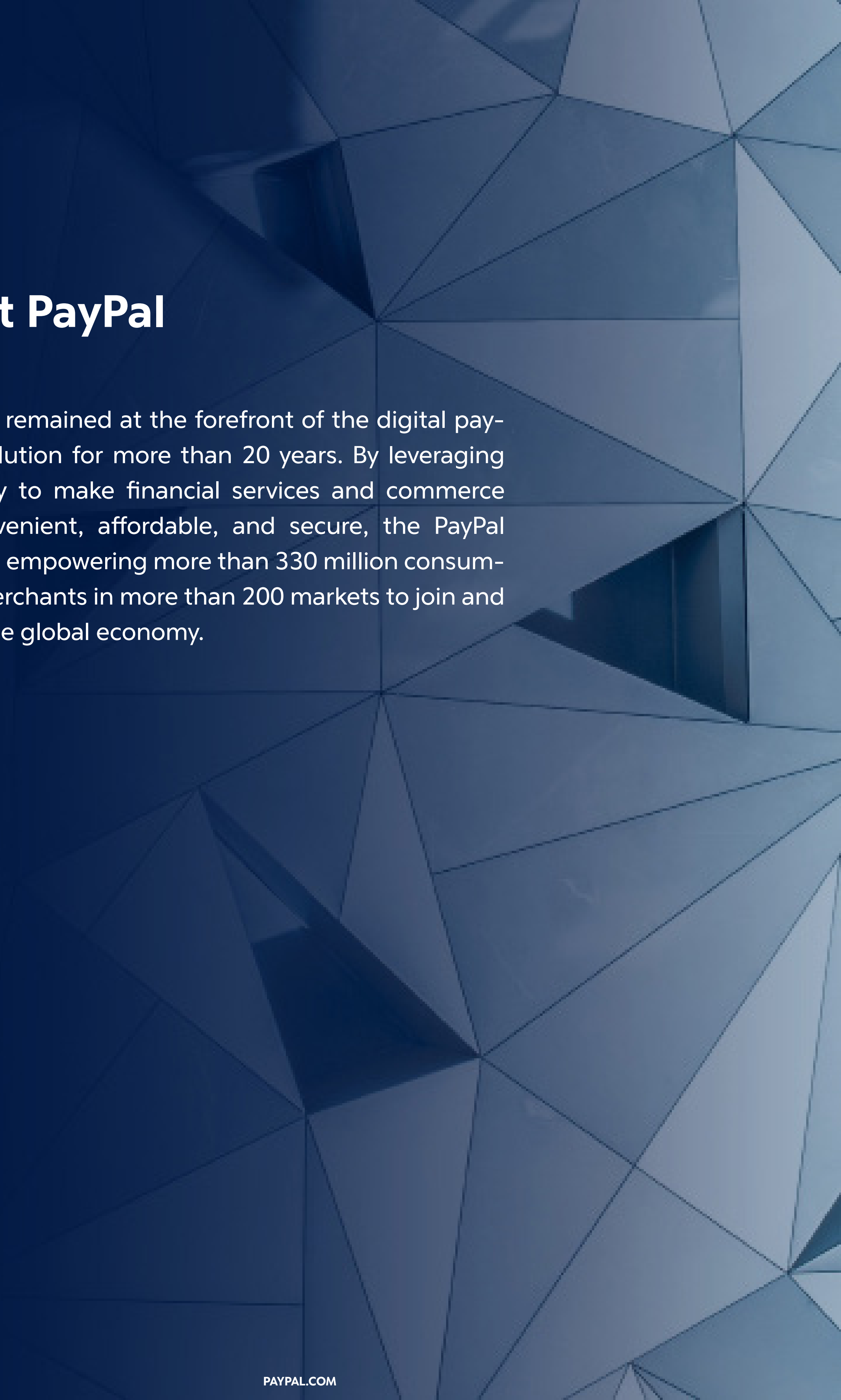
In a machine learning model, we tune a function to map input data to output data. The mapping function will only be as good as the data we provide it to learn from. There needs to be enough data to reasonably capture the relationships that may exist between both input and output features. Machine learning is a process of induction, and a model can only capture what it has seen. If your training data does not include edge cases, they will very likely not be supported by the model. For example, when we want to train machine learning models to recognize fraudulent behavior among all transactions, we will need information from both good and bad events from many transactions to create an advanced fraud detection machine learning model.

Seasonality

The amount of data also highly depends on the time frame of historical data. If there is seasonality in the data set, we will need a large enough amount of data to see a full cycle of a pattern. A machine learning model is designed to describe the way data changes; therefore, we need enough data points to capture these changes. Having a short time frame significantly complicates the development and verification of models intended to produce seasonal behaviors. What may be irrelevant when we have a lengthy time frame can be very important for short data samples. The historical data we are using for modeling should represent the actual traffic flow in real-time. For example, having one million data records from the last three months is more helpful in detecting meaningful patterns than just having on week's worth of records. The rate of data depends on the use case(s); however, as previously mentioned, historical data should be enough to capture the full cycle of fraudulent patterns.

Conclusion

Machine Learning requires a large amount of data to train algorithms. The more data you have, the more advanced machine learning model you can develop. The availability of adequate data accelerates the process of training and identifying more meaningful patterns, and as a result, a model will be more accurate and robust for further predictions.



About PayPal

PayPal has remained at the forefront of the digital payment revolution for more than 20 years. By leveraging technology to make financial services and commerce more convenient, affordable, and secure, the PayPal platform is empowering more than 330 million consumers and merchants in more than 200 markets to join and thrive in the global economy.