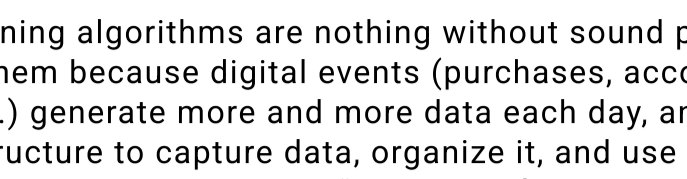


The Dummy Handbook on Data Science for Fraud Detection

This document provides an introduction to the work of data scientists in the context of using machine learning to fight fraud in digital businesses, including financial services, ecommerce, and social networks. It discusses the main activities, goals, and challenges at play in the intersection of data science and fraud detection with the main purpose of facilitating communication between data scientists and engineers and the other members of fraud prevention teams.

Ready? Here we go!



What is the Application of Data Science in Fraud Detection?

Data science is the combination of domain expertise, quantitative methods, and analytical skills to extract meaningful insights from large quantities of information. In the world of fraud detection, data science refers to the activities performed to process the data available for financial services, ecommerce, and other firms to make it as useful as possible to detect and prevent fraud.

The Importance of Data Science for Fraud Detection

You probably already know that machine learning is a game-changer for modern fraud prevention strategies, and we promise that this is not a buzzword. Without the use of artificial intelligence, fraud teams would be outmatched by organized fraudsters that employ expensive and advanced tools to commit digital crimes.

But machine learning algorithms are nothing without sound practices of data science behind them because digital events (purchases, account openings, transactions, etc.) generate more and more data each day, and without the adequate infrastructure to capture data, organize it, and use it in scalable ways, fraud teams can become inundated in "garbage information" and internal confusion can bring chaos.

If you want a primer on machine learning, or just want to refresh your memory, we recommend that you check out [The Dummy Handbook on Machine Learning Fraud Detection](#).

In the context of machine learning, the term "data engineering" is often used in a more specific sense to refer to the construction and transformation of information records (aka data sets) in preparation for use in an ML model.

Data engineering is crucial to machine learning for several reasons, but these two can very well be the most relevant:

- GIGO (garbage in, garbage out), a common tenet in computer science, holds that the quality of output is determined by the quality of the input.

In machine learning, the output is the result of the algorithm, and the input is the data that is fed to it. If the data is garbage, then no matter how sophisticated an algorithm you design, the output will sadly never amount to much more than that too. 🧐

- That time equals money might sound like a truism, but data science teams who overlook it can pay dearly.

According to IBM, in most companies, the so-called "80/20 rule" applies because 80 percent of a data scientist's valuable time is spent simply finding, cleansing, and organizing data, leaving only 20 percent to actually perform analysis. 🧐

With the mean salary for a data scientist rising close to \$130,000 per year in states like California and New York, we invite you to do the math about all the money that your team invests in data engineering processes. 🧐

Data Set Construction: The First Step for Fraud Detection

In data science, a feature is a specific property or a characteristic of the process under study. In layman's terms, a feature is the column of a datasheet that organizes important information.

Constructing a data set typically involves selecting the information that will be used in a machine learning model and integrating it into a single and usable data set. If we compare data engineering to building a house, then this stage would be like choosing the raw materials (bricks, concrete, glass, pipes, etc.) and bringing them into the site for future use.

Let's start creating our own example. If we were putting together a data set about the transaction history of an online retailer, we would start with something like this:

Transaction ID	Cardholder Name	Cardholder Address	Shipping Address	Timestamp	Type of Card
1	George Washington	450 E Street, NW Washington, DC 20001	450 E Street, NW Washington, DC 20001	1609459200	Visa
2	John Adams	1600 Penn Avenue, NW Washington, DC 22005	1600 Penn Avenue, NW Washington, DC 22005	1612137600	Mastercard
3	Thomas Jefferson	23 10th St, San Francisco, CA 94103	456 Clear Road, Houston, TX 77099	1614556800	Visa
(...)					

Columns 2 through 6 contain the features of this data set, which are each attributed to transactions identified with an arbitrary ID (column 1).

The data set construction process that yielded this table included feature selection activities where a data scientist picked the most relevant attributes out of a larger data set and imported them to create a new table.

It also included data joining activities. There are 6 types of joins: inner, left inner, left outer, right inner, right outer and outer, 🧐 and in the most common joins, two data sets are combined in a side by side manner to convert them into a single one. For this to work out, both initial data sets need to share at least one column. In other instances data sets are joined by putting one on top of each other, thus requiring all of the columns to be the same.

- 🧐 [The Data School: What Are Data Joins?](#)

Data Set Transformation: The Second Step for Fraud Detection

Transforming a data set refers to the process of modifying data to make it valuable for the specific purposes at hand. Not all data sets are created equal and, most importantly, no two use cases are identical. It is paramount to transform data into its most usable form to ensure accuracy of the ultimate machine learning application.

The processes of data set transformation are also referred to as feature engineering because logical operations and other functions are applied to features in order to transform the data set with the goal of making it as useful as possible for the specific project at hand.

Feature engineering (a.k.a. feature building or signal creation) is truly part art and part science because it is the perfect intersection of technical expertise and domain knowledge. Without a thorough understanding of the problems that the algorithm is trying to solve and all the factors that play a part in them, it would be impossible to know what questions to ask the data.

Let's continue with our example:

Through **feature creation**, we could derive new information from our data set, potentially garnering a higher predictive power than the raw data.

Txs ID	Cardholder Name	Cardholder Address	Shipping Address	Address Match	Timestamp	Type of Card
1	George Washington	450 E Street, NW Washington, DC 20001	450 E Street, NW Washington, DC 20001	Yes	1609459200	Visa
2	John Adams	1600 Penn Avenue, NW Washington, DC 22005	1600 Penn Avenue, NW Washington, DC 22005	Yes	1612137600	Mastercard
3	Thomas Jefferson	23 10th St, San Francisco, CA 94103	456 Clear Road, Houston, TX 77099	No	1614556800	Visa
(...)						

By adding a new feature named "Address Match", our data now tells us whether or not the cardholder address and the shipping address matches for each transaction, which can be a useful feature for determining if a specific event needs to be reviewed in more detail.

Through **feature transformations**, we could make data easier to understand in human terms and avoid future errors.

Txs ID	Cardholder Name	Cardholder Address	Shipping Address	Address Match	Purchase Date	Type of Card
1	George Washington	450 E Street, NW Washington, DC 20001	450 E Street, NW Washington, DC 20001	Yes	1/1/2021	Visa
2	John Adams	1600 Penn Avenue, NW Washington, DC 22005	1600 Penn Avenue, NW Washington, DC 22005	Yes	2/1/2021	Mastercard
3	Thomas Jefferson	23 10th St, San Francisco, CA 94103	456 Clear Road, Houston, TX 77099	No	3/1/2021	Visa
(...)						

In this case, what we did was use a simple function to convert the timestamp, which was expressed in Unix time (a system for describing a point in time), into a simple date format. We didn't create data, only transformed it.

Through **feature extraction**, we could automatically create new variables by obtaining them from raw data. This can give us new information in a simplified form or reduce the volume of data for manageability. There are many feature extraction methods such as text analysis, cluster analysis, edge detection algorithms, and principal components analysis. Here's an example of the first one:

Txs ID	Cardholder Name	Cardholder Address	Zip Code	Shipping Address	Address Match	Purchase Date	Type of Card
1	George Washington	450 E Street, NW Washington, DC 20001	20001	450 E Street, NW Washington, DC 20001	Yes	1/1/2021	Visa
2	John Adams	1600 Penn Avenue, NW Washington, DC 22005	22005	1600 Penn Avenue, NW Washington, DC 22005	Yes	2/1/2021	Mastercard
3	Thomas Jefferson	23 10th St, San Francisco, CA 94103	94103	456 Clear Road, Houston, TX 77099	No	3/1/2021	Visa
(...)							

In a much-simplified exercise of **feature selection**, we could transform our data set to remove certain features that are deemed less important, and focus on the most relevant ones. Perhaps certain features, such as the type of card, can be omitted for specific purposes. Feature selection can also remove redundancies and assign relative weights to different features.

Txs ID	Cardholder Name	Zip Code	Shipping Address	Address Match	Purchase Date
1	George Washington	20001	450 E Street, NW Washington, DC 20001	Yes	1/1/2021
2	John Adams	22005	1600 Penn Avenue, NW Washington, DC 22005	Yes	2/1/2021
3	Thomas Jefferson	94103	456 Clear Road, Houston, TX 77099	No	3/1/2021
(...)					

Velocity Features - What Makes or Breaks a Good Fraud Strategy

Up to this point, we have discussed the creation and transformation of certain basic features; however, there are certain more advanced features that are important to discuss. This is especially the case with velocity features, which can be defined as the result of counting events or attributes within a specified time frame.

Velocity features are created based on other features of existing data sets and can yield immensely valuable information for fraud detection because they take into account the critical dimension of time in the events that comprise the dataset. Time is a complex and inherently dynamic concept, and velocity features are data scientists' way of incorporating its effects into data sets to extract the most valuable information possible.


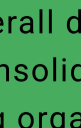
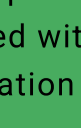

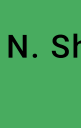
Like many other features, velocity features can be understood as the answer to "asking" questions to the dataset, and here are some examples of velocity features and how they are used in fraud prevention:

Underlying Question	Feature Name	Parameters	Interpretation
How many times has each user installed this application in the last 30 days?	Num_Apps_per_User_30days	if >3	then, bust-out fraud alert
Have unusually large sums of money been withdrawn from accounts after long dormant periods?	Savings_Withdrawn_Following_1yInDormant	if >\$500 after 1year	then, ATO alert
What is the email pattern score for all new accounts on the same IP?	Email_Pattern_Score_Aggregated_perIP	if >0.9	then, mass registration alert

Advanced Feature Engineering Techniques for Machine Learning

Did the simplified examples used above spark your curiosity for the more nuanced activities that data scientists in your team perform? Great! Mission accomplished.

Here are a few more advanced techniques used in feature engineering for ML:

 Imputation	A technique aimed at handling missing values that can impact the performance of ML models and sometimes arise due to human errors, data flow interruptions, privacy issues, and other factors.
 Outlier management	A method used to produce more accurate data representation by removing outliers from a data set. Depending on the case at hand, outliers can be removed, replaced, capped, or discretized.
 Log transformation	A technique used to turn a skewed distribution into a normal or less-skewed one by taking the log of the values in a column and using them in a new one to handle confusing data and transforming it to normal applications.
 One-hot encoding	Since most ML algorithms can only ingest numerical data, this technique is used to convert categorical features (e.g. colors or names) into numbers. Instead of using a column to describe the color of a house, you could have a column for each color available and a yes/no answer for each house in the form of 1/0.
 Scaling	A tool to allow machine learning algorithms to make sense of data that comes in scales that are very different among each other through normalization and standardization operations.

Source: Harshil Patel - Towards Data Science, 2021

Real-Life Examples Involving Data Engineering

Data science teams at Google Maps have in all likelihood invested millions of hours and dollars into creating and transforming the data sets that power their route calculation and live driving direction applications. Can you imagine how much data Google has at its disposal for this? But making sure that the right features and only the right features are fed to models must be a huge task. [Here's a little more detail into how they do it.](#)

Have you ever wondered how facial recognition software, such as the one found in modern iPhones, works? Well, there's a machine learning algorithm behind it, and behind this algorithm there is a lot of work by data scientists that engineered the database that feeds it. [Here's an in-depth explanation of the features it uses and the way these features were engineered.](#)

Main Feature Engineering Needs for Modern Fraud Teams

Back to our favorite topic, fraud prevention. We've covered why data science is so important for fraud teams and what data scientists actually do when they are performing data engineering. Now it's time to talk about the challenges they face and the tools that they need to be successful in their jobs.

The top three challenges in feature engineering for machine learning fraud prevention are:

- ➊ **Speed to value.** Long development and deployment cycles for new features and models often translate to huge losses before new and emerging fraud patterns are detected and blocked.
- ➋ **Minimizing IT dependency.** Feature engineering processes must be performed without increasing reliance on already saturated resources in information technology departments.
- ➌ **Agility with data.** Being able to bring in new data sources for features and models without interrupting or disrupting their operation is a make-it-or-break it capability in modern fraud detection strategies.

The Feature Platform Difference:

Criteria	DataVisor	Competition
Real time join and computation	✔️	Often requires time consuming data preprocessing by IT
Flexible sliding window for aggregation	✔️	Rigid configuration for long time series aggregation, relies on offline join and restrictive time-window in real time
Feature freshness	✔️	Features refreshed on daily basis or even longer
Fast feature retrieval	✔️	Reliance on SQL queries for long time series aggregation in real time often result in poor performance
Multidimensional and complex features	✔️	Key-value pair based feature design does not support complex multidimensional features well
Fast backtest and backfill	✔️	Dependency on IT to solve efficiency issues lead to higher cost for data integration and feature computation

Easy Integration of Multiple Data Sources with Feature Platform



Machine Learning and Data Science for Fraud Lingo:

- **Big data** - A reference to data sets that are too large or intricate to be processed and made sense of with traditional means.
- **Clustering** - An unsupervised machine learning technique that classifies each observation in a data set into specific categories, known as clusters, according to their shared properties.
- **Data engineering** - The construction and transformation of information records (aka data sets) in preparation for use in an ML model.
- **Data join** - The acts of merging two or more data sets into one in a cohesive way that allows usability.
- **Data mining** - The use of computers to analyze large data sets to look for patterns.
- **Data science** - The study and practice of extracting knowledge and insights from large and complex sets of data.
- **Data Wrangling** - A term that is often interchanged with data cleaning, data remediation, and data munging and refers to the processes performed to transform raw data into more manageable formats.
- **Feature creation** - The process of generating attributes in a dataset that were not previously represented, often done by applying mathematical or logical operations on existing features.
- **Feature extraction** - The process in which a data set is divided and the number of attributes in it is reduced into a more manageable amount.
- **Feature selection** - The process of singling out the attributes of a data set that are the most relevant to the process at hand and setting aside the ones that are not deemed relevant.
- **Feature transformation** - The princess of modifying the attributes of a dataset to make them more usable for the purpose at hand, often by manipulating the way they are expressed.
- **Feature** - In ML-speak, the expression for a piece of measurable information about an event or element. The expression is often interchanged with attribute, property, and field.

References and Further Reading:

- [Towards Data Science](#)
- [Data Science Glossary](#)
- [Discover Feature Engineering, How to Engineer Features and How to Get Good at It](#)
- [What Is Feature Engineering – Importance, Tools and Techniques for Machine Learning](#)
- [Heavy AI - Feature Engineering](#)
- [Harvard Business School - DATA WRANGLING: WHAT IT IS & WHY IT'S IMPORTANT](#)

Are you curious about how you can improve the account security at your organization? A fraud specialist can answer any questions you may have in a no-pressure environment.

[Schedule a free consultation session here!](#)

About DataVisor

DataVisor is the world's leading AI-powered Fraud and Risk Platform that delivers the best overall detection coverage in industry. With an open SaaS platform that supports easy consolidation and enrichment of any data, DataVisor's solution scales infinitely, enabling organizations to act on fast-evolving fraud and money laundering activities as they happen in real time. Its patented unsupervised machine learning technology, combined with its advanced device intelligence, powerful decision engine and investigation tools, provides guaranteed performance lift from day one.

For more information on DataVisor:

- ✉ info@datavisor.com
- 🌐 www.datavisor.com
- 📍 967 N. Shoreline Blvd. | Mountain View | CA 94043