GenRocket

MACHINE LEARNING CASE STUDY

GenRocket's Synthetic Data Accurately Trains AI-Assisted Tax Fraud Detection System

Tax fraud is present in every category of tax collection including personal and corporate income taxes, social security contributions, sales taxes in the US, and value added taxes (VAT) collected in more than 170 countries. **In our increasingly digital and global economy, VAT fraud and sales tax evasion has become a major source of lost revenue for governments around the world.**

According to a 2018 report by the European Union's Anti-Fraud Office (OLAF), VAT fraud in the EU is estimated to cost between €40 billion and €60 billion each year. This represents a significant loss of revenue for governments that could be used to fund public services and infrastructure.

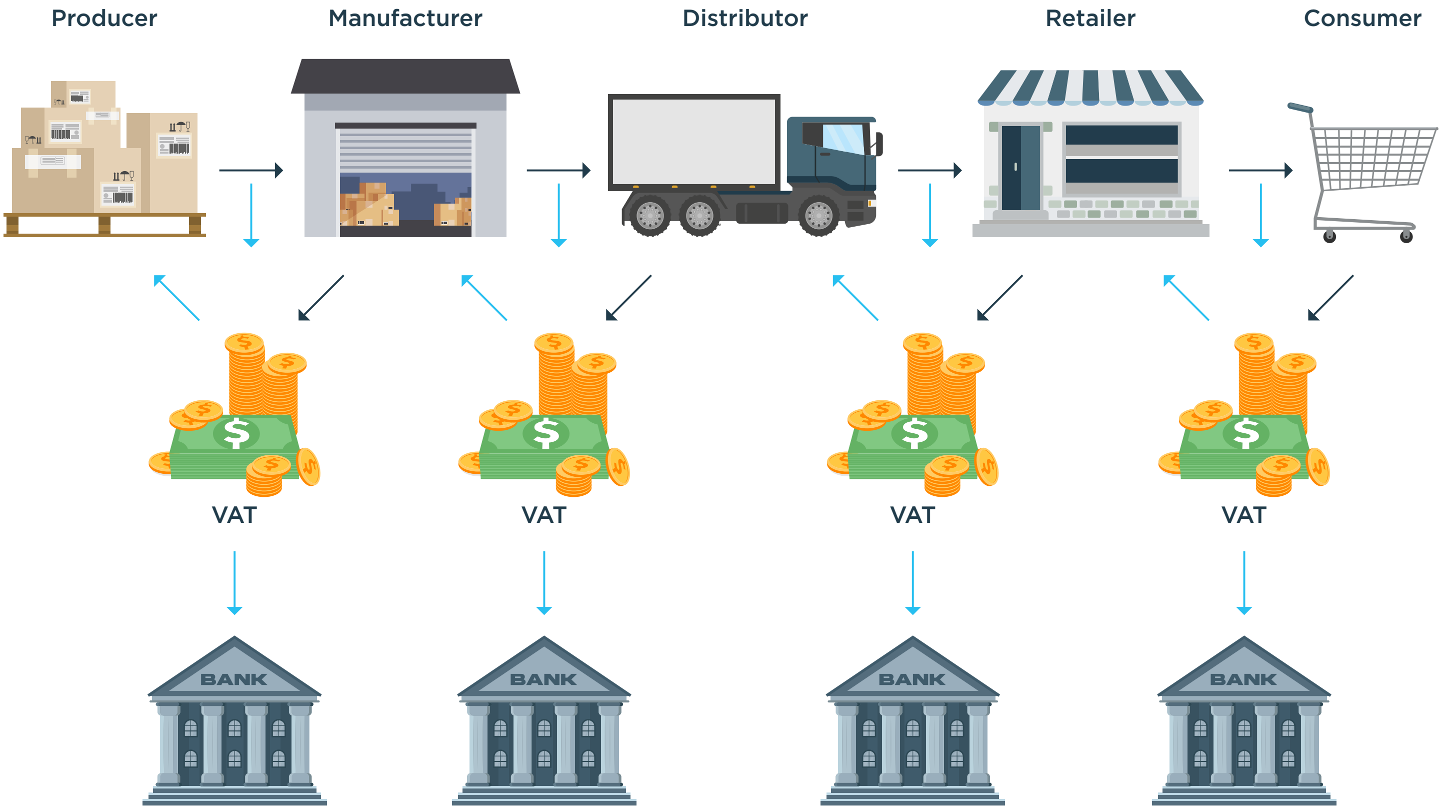
In the United States, sales tax fraud is estimated to result in billions of dollars of lost revenue each year. A 2019 study by the National Conference of State Legislatures (NCSL) estimated that the U.S. states collectively lose around $20 billion per year in uncollected sales tax revenue.

Catching Tax Fraud with the Help of AI

VAT fraud involves businesses or individuals failing to properly collect and remit VAT to the appropriate government authorities. It's difficult to detect in the massive volume of commercial transactions continuously flowing through complex, multinational supply chains. Each step in the supply chain may involve a different tax calculation depending on the tax laws where those transactions take place. VAT fraud is currently being committed by producers, wholesale distributors, retailers, and consumers. And unfortunately, much of the abuse goes undetected.

To help governments in more than 160 countries detect tax fraud, a major software company developed artificial intelligence technology for a sophisticated tax fraud detection system. Tax fraud detection is an area where machine learning and artificial intelligence can excel when properly tested and trained by synthetic data.

VAT Fraud in a Global Supply Chain



The architects and software developers at this company knew they needed large volumes of data to properly train their machine learning algorithms. And they knew they needed training data with **exact precision** – for both normal transaction values as well as for outliers or anomalies. This is essential for ensuring the system will consistently recognize valid tax calculations, while at the same time, detect any occurrences of erroneous or missing tax amounts.

The company was in a "greenfield" situation where there was no production database to copy, so they needed a synthetic data generation platform that could generate machine learning training data from a detailed, written specification. And to eliminate bias and data quality issues, they needed a synthetic data platform that has precise control over the data generation profile so that highly statistically accurate training data can be produced, in huge quantities.

Most synthetic data tools build machine learning training data by discovering the statistical distribution profiles of data in a production database. These tools generate a replica of the production database as a synthetic data copy that has a matching statistical profile. The result is secure synthetic data that can be used as training data. **However, a common problem with a "synthetic replica" of a production database is that it reproduces the biases and data quality issues (e.g., corrupt, inconsistent, or missing data).** And a synthetic replica almost always under represents the outliers (anomalies) that can represent fraudulent behavior.

The ability to control statistical profiles and data variations in a training dataset improves the accuracy of the model and minimizes false positives and false negatives. **For training data to serve as the "programming language" for AI and ML, there must be precise control over the volume and variety of data generated for training the algorithms for any given ML use case.**

Testing a Complex Tax Fraud Detection Algorithm

To build the ruleset for training their machine learning models, the software company's team of programmers and financial experts started by studying the tax rules for a mid-sized country. The project team created an 18-page specification that accurately defined the rules and data scenarios needed to train and test their tax fraud detection system.

The spec required 1,440,090 organizations and 990 million invoices to cover the numerous tax provisions applicable to all participants of their supply chain scenarios. And because of their pressing deadlines, the software company needed the project setup and all the data generated *in less than 30 days.*

Many complex statistical characteristics associated with this economy needed to be modeled. For example, the organizations needed to be generated in the following distribution:

Enterprise Classification	Size Composition
Micro Organization	83.87%
Small Organization	12.30%
Medium Organization	3.01%
Large Organization	0.70%
Big Organization	0.16%

Many attributes of these customer organizations needed to be modeled in addition, such as whether their business was local or international, the age and year of registration, and a table of 90 supply chain transaction activities needed to be mapped to the organization type. Additionally, they needed a defined distribution of invoice volume with a minimum and average value per organization size. Then layered on top of this transaction data profile was the need to model the tax laws with rules and thresholds for defining valid and invalid tax payments.

GenRocket Selected as the Tool of Choice for Machine Learning Training Data

The software company approached numerous synthetic data providers with their requirements but only found one synthetic data platform that could generate the volume and accuracy of synthetic data based on a written specification. GenRocket was selected as the ideal vendor for this machine learning data requirement and delivered the project, as required, in less than 30 days.

The project validated the quality, quantity, and accuracy of the synthetic data produced by GenRocket and its ability to accurately train and test complex machine learning models.

The Future of Machine Learning Data Is GenRocket Synthetic Test Data

With the explosive growth of machine learning and AI-based systems, developers need synthetic training data solutions that provide the best fit for their application requirements. **Rules-based synthetic training data can be an extremely effective choice for anomaly detection when there is expert knowledge of the data environment and anomalies can be defined by rules or thresholds.** With rules-based synthetic training data, developers can effectively design the data to simulate a near-infinite variety of data distributions that simulate both normal and anomalous conditions. **The use of controlled and conditioned synthetic data can more accurately train the anomaly detection algorithms to discriminate between valid and invalid events.**

AI-assisted anomaly detection often involves multiple systems, with complex interactions, and rulesets for which production data, or a synthetic replica, is lacking. **A wide range of applications in financial services, healthcare, cybersecurity, manufacturing, and many others can benefit from rules-based synthetic training data.**

GenRocket proved with this major software company that it could meet the challenge of generating massive volumes of synthetic data with full control over the variety needed to accurately train a complex set of machine learning algorithms. And with GenRocket, the customer’s tax fraud detection system can be iteratively trained to keep up with changes in tax laws, catch new methods of tax evasion, and be used to generate training data for additional countries, economies, and tax authorities.