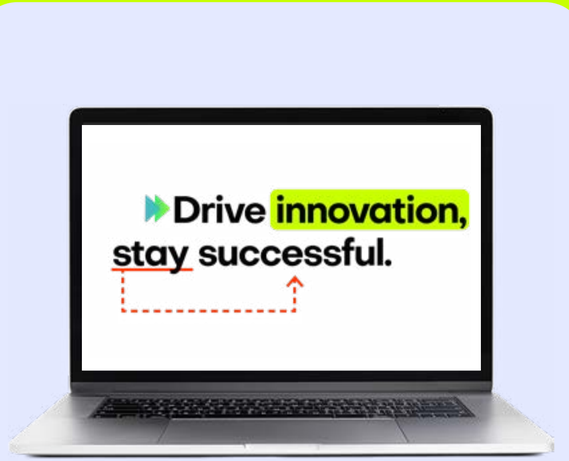
Case study

MOSTLY AI

Power Up Fraud Detection Models with Synthetic Data



Fraud detection is a complex problem with many cutting-edge AI/ML solutions.

Challenges

Fraud detection is a complex problem with many cutting-edge AI/ML solutions. However, these algorithms are only as good as the data used to train them. Traditionally used rule-based systems produce a high number of false positives and a labor-intensive follow-up process. Investigating a single customer for potential fraud can cost up to $24,000.[1] AI/ML algorithms help reduce false positives and detect new frauds, but their performance is highly dependent on the quality of the training data. Rare, high-value frauds are often missed, and signals alerting to fraudulent activity can be misleading.

Solution

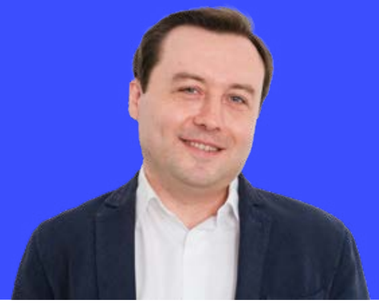
Upsample fraud patterns with synthetic training data to boost machine learning performance. Synthetic training sets are better than real data because by balancing the dataset, the model is able to detect fraud cases more efficiently, resulting in a consistently high AUC number. The “Area Under the Curve” (AUC) metric is used to evaluate how well a binary machine learning classifier (like a fraud detection algorithm) is performing. It is calculated by measuring the true positive rate against the false-positive rate. In other words, it explains how good the model is at separating fraud and nonfraud cases. The aim is to have a consistent and high AUC so that any threshold selected will have a high true-positive rate and low false-positive rate. Fresh, large batches of synthesized training data should be generated from raw transactional datasets periodically to recalibrate the algorithm and catch new patterns and signals.

1 How financial firms help catch crooks, The Economist https://www.economist.com/the-economist-explains/2017/11/28/how-financial-firms-help-catch-crooks

Result

Through various case studies, MOSTLY AI's synthetic data platform has shown to have a consistent improvement on the AUC curve from relative 2–15% compared to using raw, imbalanced data.[2] An improvement of 10% could yield a 10% decrease in false positives. Consider a model with a false positive rate of 1%. The model has identified 100 000 positive fraud cases, out of which ~1000 might not actually be fraud. If we lower the false positive rate to 0.9%, such that only ~900 are not correctly identified. Having ~100 fewer cases to investigate could lead to a savings of $2.4 Million.

2 Boost your Machine Learning Accuracy with Synthetic Data, Michael Platzer https://mostly.ai/2020/08/07/boost-machine-learning-accuracy-with-synthetic-data/



Upsample fraud patterns with synthetic training data to boost machine learning performance.

Highlights

10%

improvement in machine learning performance

100

fewer cases to investigate

$2,4M

saved by reducing false positives

About MOSTLY AI

MOSTLY AI is the leading synthetic data company globally. Its platform enables enterprises across industries to unlock, share, fix and simulate data.

Thanks to the advances in artificial intelligence MOSTLY AI's synthetic data looks and feels just like real data, is able to retain the valuable, granular-level information, yet guarantee that no individual is ever getting exposed. This enables businesses to drive innovation and digital transformation, overcome data silos, improve machine learning models as well as application testing capabilities. MOSTLY AI was founded in 2017 and is headquartered in Vienna, Austria. Its global operation includes customers in a variety of verticals, including banking, insurance and telecommunications.

Contact: hello@mostly.ai

Vienna office (HQ)
MOSTLY AI Solutions MP GmbH
Hegelgasse 21/3 · 1010 Vienna · Austria

New York office
MOSTLY AI Inc.
500 7th Ave 8th floor · New York · NY 10018 · United States

▷ Talk to one of our experts

MOSTLY·AI