clearbox AI

humans && machines

SYNTHETIC DATA FOR DATA AUGMENTATION

Improving Fraud Detection models with Synthetic Data

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai

Introduction

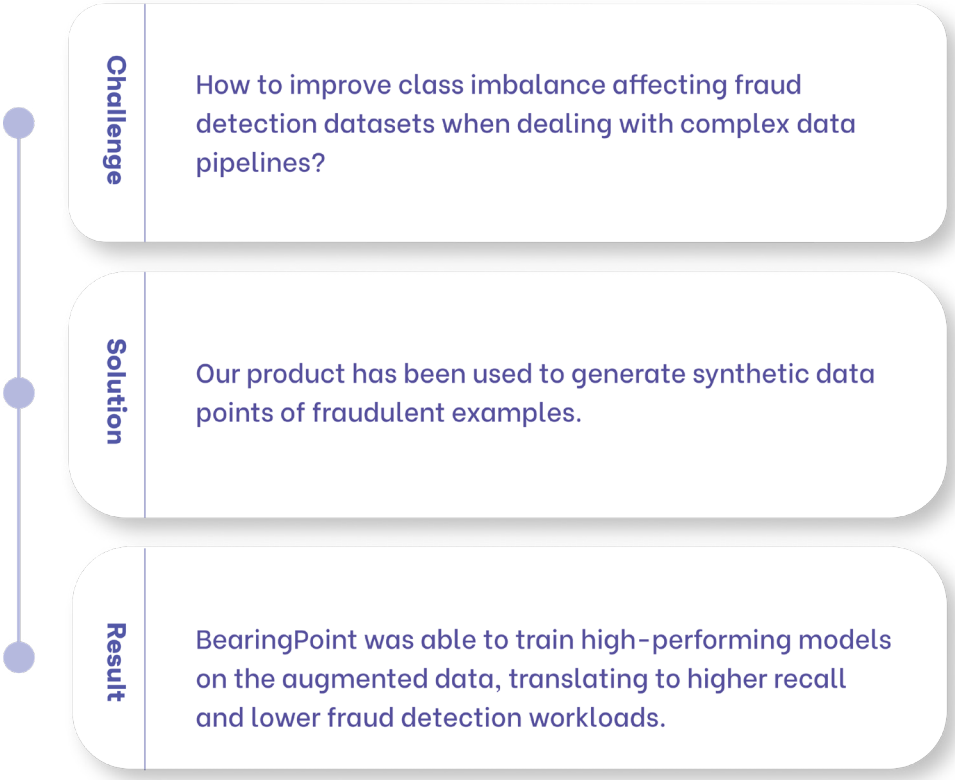
Fraud detection is a critical task in the financial domain, and machine learning is often regarded as a promising way to automate it. Machine learning is usually used to flag suspicious requests that need to be manually checked by a domain expert. Reducing the number of flags will strongly reduce the workload of the domain expert, at the same time it is paramount that no fraudulent case goes unflagged.

Unfortunately, datasets used for training and evaluating these algorithms can be affected by a **strong class imbalance**, which occurs when the number of examples of one class (e.g. fraudulent transactions) is significantly lower than the number of examples of the other class (e.g. non-fraudulent transactions). This issue makes it difficult to train robust and performing models, and it usually corresponds to a higher workload of domain experts who have to check each flagged request manually.

For this use case, the most important metrics to monitor in order to assess the efficiency of the automation process are the model's Precision and Recall. In particular, the recall measures the ability of a model to identify all fraudulent requests correctly.

BearingPoint is an independent management and technology consultancy with European roots and a global reach. The company is one of the leading providers of ML-based fraud detection models and was exposed to these issues daily. Thanks to the test on the field that we carried out with BearingPoint, we demonstrated how, through the generation of synthetic data, it is possible to increase

model performances while decreasing the number of transactions that must be manually checked.
In particular, Clearbox AI's Enterprise Solution helped them tackle class imbalance problems while working with one of their clients who commissioned them a fraud detection model.



Challenge

Several techniques can be adopted to improve class imbalance. **Oversampling**, for example, consists in creating synthetic minority examples to re-balance the original dataset. SMOTE is one of the most popular techniques which has been proven to be useful in many applications. The problem arises when the cardinality of the dataset increases. This is often the case for fraud detection use cases where we want to make use of as much information as possible. In this case the synthetic examples generated by SMOTE start becoming more and more unrealistic. It is therefore necessary to use alternative methods, for example based on **generative models**.

Solution

BearingPoint installed our Enterprise Solution on the infrastructure of one of their clients, a retail bank. They connected it to a relational database containing transaction histories and used our tool to quantify class imbalance and find the best data augmentation strategy. They finally generated an **enriched dataset** containing the original clients plus several synthetic fraudulent examples. They used this dataset to train a machine learning model based on boosted trees.

Result

Accessing the augmented dataset allowed BearingPoint to considerably **reduce the number of false flags**. As we demonstrated, the model trained on augmented data presented a **Recall improvement of 15%** (+12% in respective of the best combination of under/oversampling). This automatically translates into more efficient and cost-effective fraud detection workflows and workloads.
These results can be extended to other use cases as well, both in the financial sector and others. The technology can be applied to any sector that needs a lot of data to improve its processes. For example, insurance, energy, telco, urban mobility, retail, and healthcare.

humans && machines

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai