clearbox AI
humans && machines

SYNTHETIC DATA FOR DATA AUGMENTATION

Improving models with synthetic training data. A project with Banca Sella.

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

VAT ID: (IT)12161430017

clearbox.ai

Introduction

<1%
original minority labels

2x
model recall

Increased
model generalization

Training models on **imbalanced datasets is always a challenging task**, especially when dealing with extreme imbalance (minority labels <1% of the data). For this use case, we have worked with the Data Science department of Banca Sella to prove how a model that runs on strongly imbalanced data can be improved by adding synthetic minority examples generated by our Enterprise Solution.

Banca Sella is an Italian financial holding. The banking company of the group is one of the largest banks in Italy by total assets.

Challenge

Imbalanced datasets make it difficult to extract useful predictive information. Traditional techniques such as SMOTE fail at generating useful minority examples.

Solution

We employed Clearbox AI's Enterprise Solution to generate synthetic minority examples to improve model training.

Result

The model trained using additional synthetic examples achieved better metrics on hold-out test data, in particular, the recall was almost doubled.

Challenge

An imbalanced dataset is an annotated dataset where the distribution of the target labels is very uneven, i.e. a label might appear much less frequently than the rest. This issue often affects datasets ranging from fraud detection to predictive maintenance, where the predictive target refers to situations that rarely happen. The usual approach is to perform oversampling, i.e. generate new synthetic minority examples to enrich the training set.

For this challenge, our Synthetic Data Enterprise Solution has been used to generate synthetic minority examples for a dataset to train a machine learning model to improve uplift metrics of one of their financial products. In this case, the dataset was characterised by the presence of an **extreme imbalance**, i.e. less than 1% of positive labels. This was a particularly challenging dataset as traditional oversampling techniques, such as SMOTE, were not able to generate synthetic examples useful for the training process.

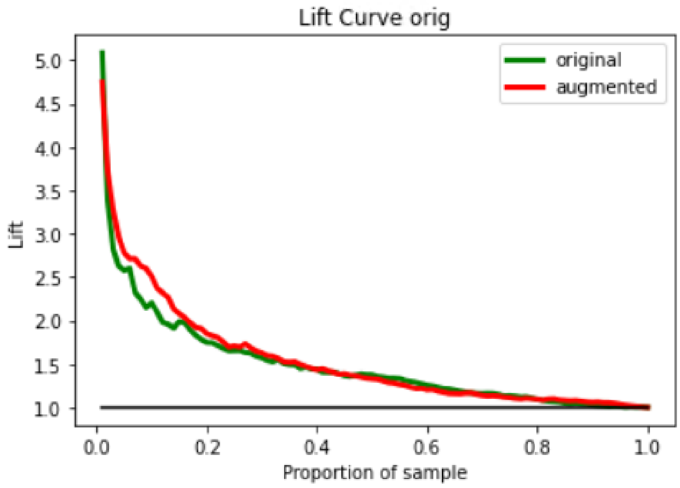
Solution

We helped the client solve this challenge through our Synthetic Data Enterprise Solution, **generating synthetic points** representing the minority examples. We segmented the original data in an unsupervised way through the **data profiling and assessment tool** while determining **anomalies** and **outliers**.

This information **facilitates the parameterisation** of the synthetic output, for example, by focusing on a particular customer type. We then used the Synthetic Data Engine to generate synthetic training points to **augment the supervised learning process**.

Result

The uplift model trained using additional minority examples was associated with better metrics when tested on holdout data. The figure below shows how the **lift curve improved when using synthetic examples**.



Regarding standard classification metrics, the most significant improvement was obtained in the **recall**: its value was **almost doubled (7% → 12%)** when using the model trained additional synthetic examples. Thanks to the additional synthetic training data the marketing department was able to make one of their most valuable tools more performant on real-world customers.

humans && machines

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

VAT ID: (IT)12161430017

clearbox.ai