SYNTHESIZED

A Guide to Data Augmentation and Data Rebalancing

REAL-WORLD APPLICATIONS OF SYNTHETIC DATA

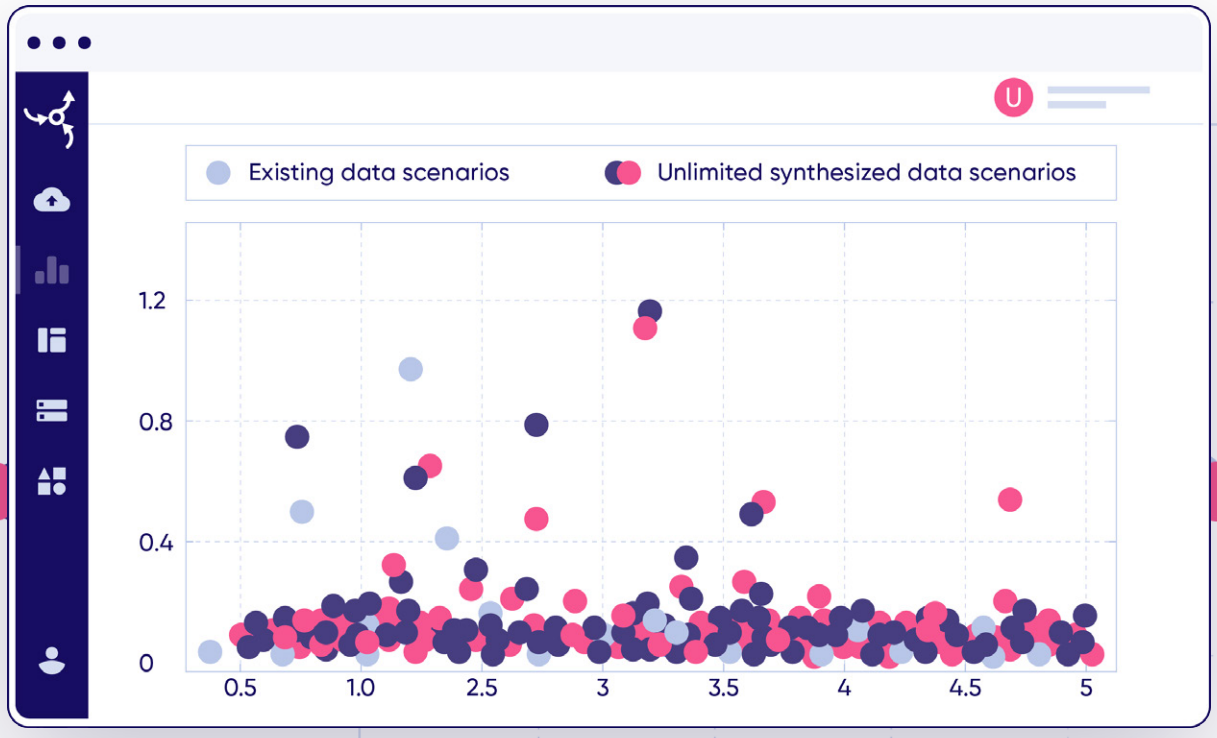


Table of Contents

Summary

This whitepaper presents some of the problems that commonly arise in Data Science and Machine Learning teams, and shows how they can be easily tackled by making use of the Synthesized DataOps platform.

Executive Summary

Data Science projects present a unique array of challenges. The Synthesized Dataops platform can help to tackle many of them, including:

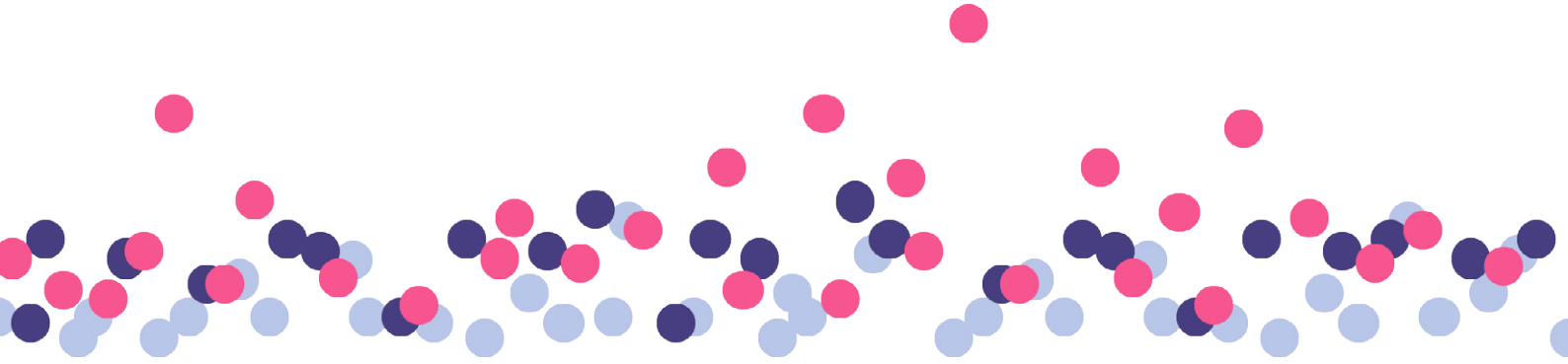
- Predictive models trained on biased datasets can result in poor performance for organisations and their customers, damage customer experience and affect brand reputation. **With the Synthesized platform, businesses can generate samples of an under-represented customer category to unbias the dataset and improve the model's performance**.
- A combination of strict compliance regulations, questions of customer trust, and the limitations of data itself can lead to data shortage. **Expanding a dataset with Synthesized improves a model's stability and performance**.
- Models can break in production if not properly validated due to population shifts. **Unseen scenarios can be generated with Synthesized data manipulation, and used to test the model's performance and ensure its stability**.

Technical Summary

Many Data Science and Machine learning projects face problems such as imbalanced, biased, and/or small datasets, as well as under/overfitting. We have used Synthesized to overcome these issues. Here are some of our most important findings:

- A predictive model trained on an imbalanced dataset can overfit on the majority class if trained to maximize the total number of correctly predicted labels.
  Synthesized provides a simple and easy way to manipulate distributions and generate a rebalanced dataset, while preserving data privacy.
- Data Augmentation is a form of regularization, as it reduces dataset variance.
- **Using Synthesized to generate new samples and appending them to the original dataset we've shown that**:
  - For small datasets, small variations and outliers affect performance strongly, making the results unstable and highly variable on each train-test split.

  - **Adding Synthesized data reduces training set variance and helps stabilize results.**
  - This regularization can also translate into better performance for large datasets. **Synthesized understands the underlying structure of the data and, by generating extra data, it can help the predictive model exploit insights not covered in the training set.**
- Some algorithms are more sensitive to population changes. **Synthesized can be used to generate different populations with distinct distributions, to validate any model and ensure its stability before moving to production.**



Introduction

Synthetic data technologies have increasingly been adopted by leading insurance companies and businesses providing consumer-facing financial services in the last three years. When implemented accurately, the same results can be obtained with Synthesized datasets [7], and the benefits of Synthesized data include full data privacy compliance and faster product development and testing. Generating high-quality data can take as little as 10 minutes for complex datasets.

But this is just the tip of the iceberg, and there are many other ways to exploit Synthesized data in data science and machine learning. In this white paper, we explore powerful applications of synthesized data in data science, compare different techniques and scenarios, and show how the data produced by the Synthesized *platform* can help in these situations.

Datasets

To showcase the business benefits of Synthesized data in data science and machine learning, we focus on six realistic datasets, see Table 1.

DATASETS	
1. **Credit Loan Default**	DESCRIPTION: Credit scoring dataset
	URL → Give Me Some Credit
	NR. ROWS: 10000
	NR. COLUMNS: 12
2. **Credit Card Fraud**	DESCRIPTION: Fraudulent/genuine credit card transactions
	URL → Credit Card Fraud Detection
	NR. ROWS: 30000
	NR. COLUMNS: 31
3. **Bank Churn**	DESCRIPTION: Predicting Churn for Bank Customers
	URL → Predicting Churn for Bank Customers
	NR. ROWS: 10000
	NR. COLUMNS: 14
4. **Telecom Churn**	DESCRIPTION: Customers who left telecom service
	URL → Telco Customer Churn
	NR. ROWS: 7043
	NR. COLUMNS: 21

DATASETS	
5. **Japanese Credit Application**	**DESCRIPTION:** Customers who left telecom service
	URL → Japanese Credit Application Dataset
	NR. ROWS: 690
	NR. COLUMNS: 16
6. **Absenteeism at Work**	**DESCRIPTION:** Absenteeism at work at a courier company in Brazil
	URL → Absenteeism at Work
	NR. ROWS: 740
	NR. COLUMNS: 21
7. **Online Shoppers Purchasing Intention**	**DESCRIPTION:** Revenue Generation from Online Customers
	URL → Online Shoppers Purchasing Intention
	NR. ROWS: 12330
	NR. COLUMNS: 18

TABLE 1.
Summary of the datasets used in this white paper.

Data Rebalancing with Applications in Personalised Marketing and Customer Segmentation

Customers always prefer to get personalized financial services that match their needs and lifestyle. Businesses offering customer-facing financial services satisfy these demands with the help of artificial intelligence and advanced data science, extracting the insights from data which encapsulates consumers' preferences, interaction, behaviour, lifestyle details and interests. The personalisation of offers, policies and pricing largely contribute to the rates of the business.

Marketing departments apply various techniques to increase the number of customers to target their marketing strategies. Customer segmentation plays a pivotal role in this process. Algorithms perform customer segmentation according to their financial sophistication, age, location, etc., classifying customers into groups by spotting similarities in their attitude, preferences, behaviour, or personal information. As a result, target cross-selling policies may be developed and personal services may be tailored for each particular segment.

A major obstacle to building and validating marketing strategies is accessing representative data about customer segments [7]. The most valuable information for the business is commonly hidden in an under-representative customer category. For example, the online shoppers purchasing intention contains 12,330 sessions, of which only 1908 (15.47%) ended in shopping, and the credit loan default dataset contains 663 (6.6%) defaulters out of 10000.

A way to overcome this issue is to generate new samples for an under-representative category, thereby rebalancing the dataset. Here, we compare three different techniques:

- **Random over-sampling.** This approach consists of adding samples from the minority class with replacement to the dataset.
- **SMOTE (Synthetic Minority Over-sampling TEchnique) [1].** This approach is based on a geometric technique in which a line is drawn between two samples of the minority class and a third sample point is created at a random location on this line.
- **Synthesized data manipulation tool.** This approach consists of learning the insights and the high-dimensional structure of data, enabling data scientists and BI analysts to automatically create new sample points that belong to the minority class.

We provide further evidence and compare the three methods. To evaluate the performance of rebalanced datasets, we use the so-called AUC score, as it is a widely used metric for imbalanced datasets.

To check how the minority class proportion affects the final results, the following procedure is applied:

- The data set is split into the training and test sets with ratio 4:1.
- The training set is resampled from the original proportion to 1:1, so that both classes have the same number of samples.
- We compute the evaluation metrics on the test set that remains unseen.

We compute the AUC metric as we resample data from the original dataset until we have the same number of samples for both classes. The results of 10 Monte Carlo simulations are shown in Figures 1. We can clearly observe an uptrend in the AUC score as the datasets are resampled. Of the three techniques compared in this study, the Synthesized data manipulation toolbox shows the best performance.

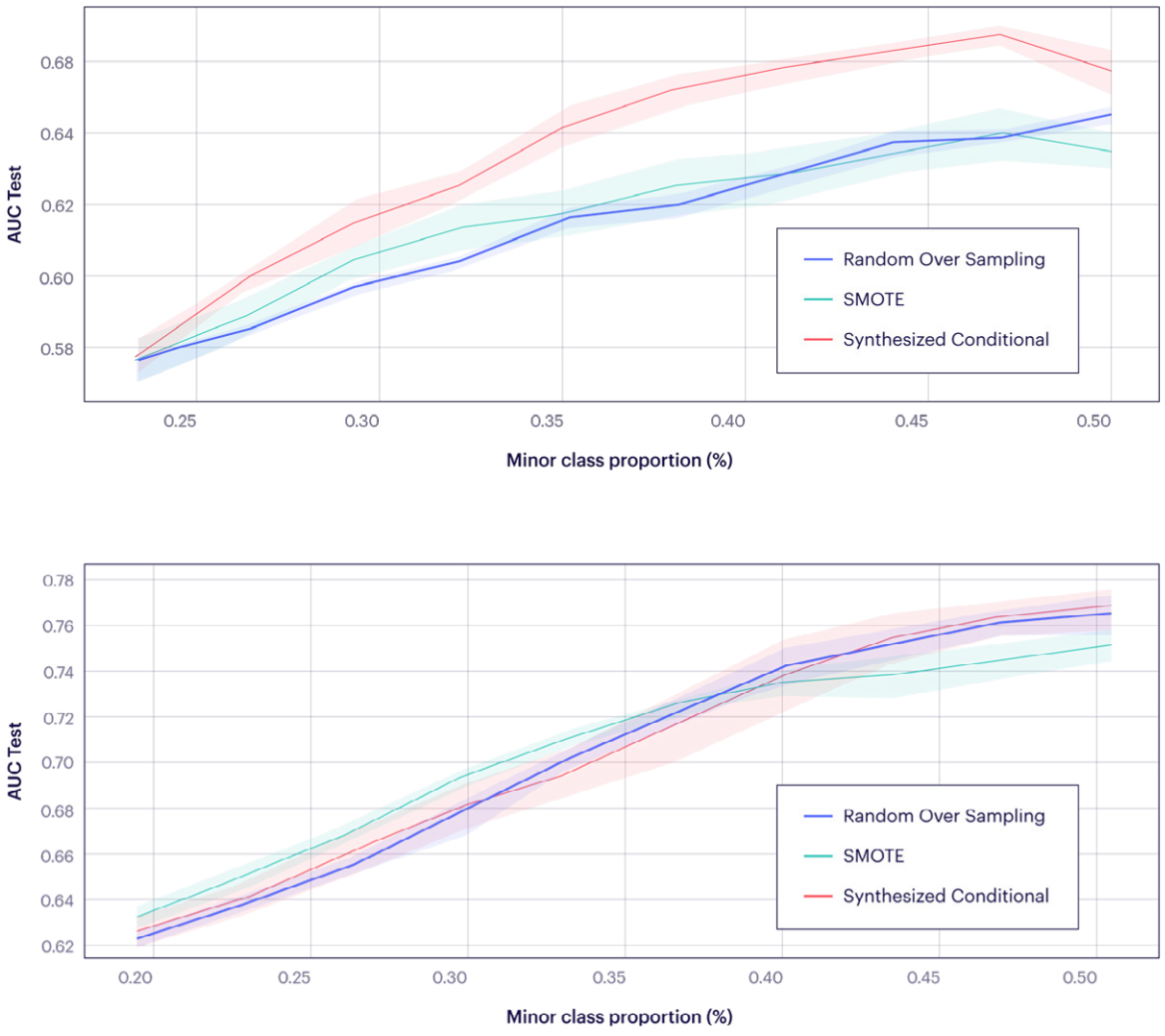


FIGURE 1.
Progression of AUC as we resample the target variable for the Bank Churn (top) and the Online shoppers purchasing (bottom) datasets.

Besides, Figure 2 shows the ROC curve and PR curve for distinct techniques for the credit loan default dataset, comparing the original with the resampled dataset. Again, the resampled techniques outperform the original dataset. In this case, all techniques exhibit similar behaviours, but Synthesized shows one key difference. **Unlike the other techniques, the privacy of the data is protected, so the data scientist can still look at the data and manipulate it without viewing any sensitive user information, as the technology used in Synthesized to generate data ensures full compliance with data privacy regulations** [8].

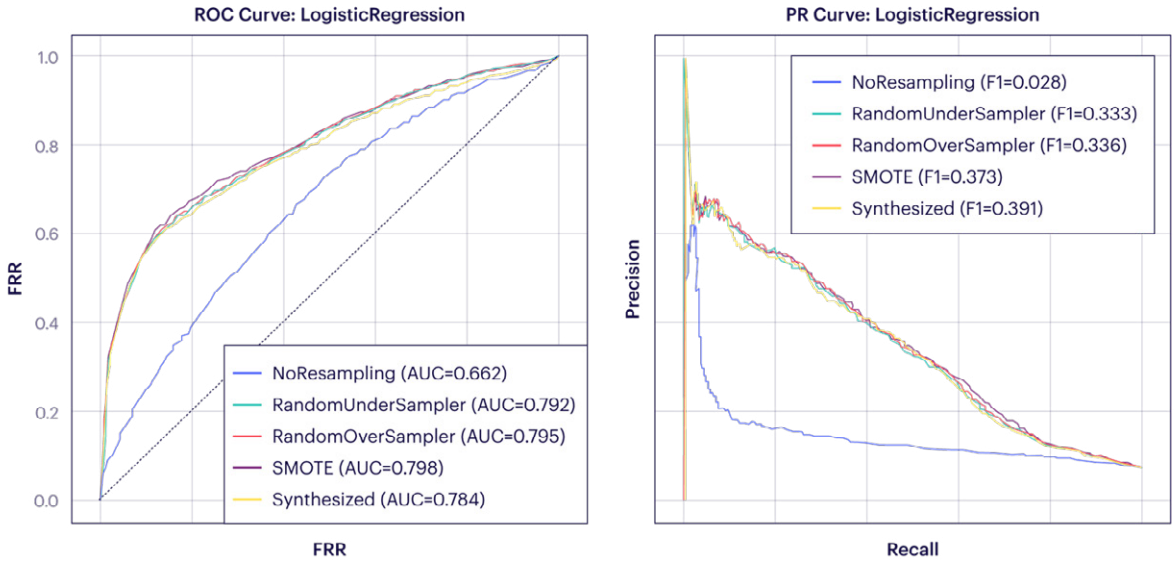


FIGURE 2.
AUC and PR curves for the credit scoring dataset, before and after resampling the dataset with different techniques.

Previously in this section, the machine learning algorithms have been evaluated at a high level, but it's critical to detect where precisely the algorithm is making wrong decisions, as the cost of a false negative can be huge compared to a false positive. Both credit scoring and online shopper purchasing datasets are good examples of this, as giving credit to a defaulter is much more costly than not giving credit to a non-defaulter, and similarly targeting a non-buyer is usually less expensive than losing a buyer.

Figure 3 sheds light on this matter, showing the confusion matrix for both datasets. A Random Forest is trained on the original (left) and re-sampled with Synthesized (right) sets. In the first case, the majority of errors are concentrated on false negatives rather than false positives, while the resampled case, the number of false negatives is drastically reduced.

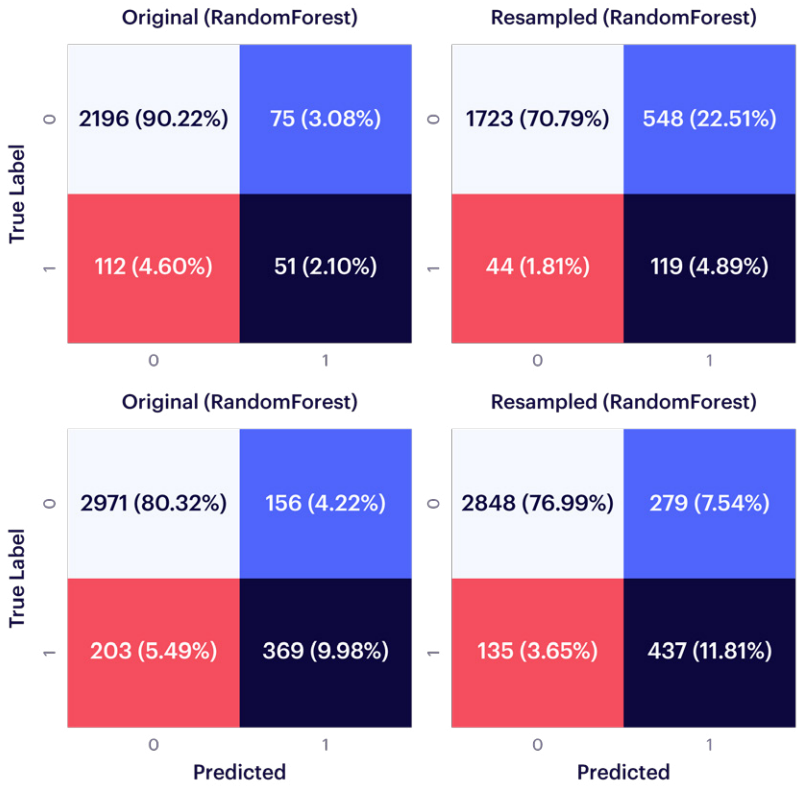


FIGURE 3.
Confusion matrix for the credit scoring (top) and online shoppers purchasing (bottom) datasets. At the left, the model (Random Forest) has been trained with the original data, and on the right, Synthesized has been used to re-sample the training set.

In summary, the experiments presented in this section showed how resampling an imbalanced dataset can heavily affect the performance of the classification model. Synthesized is simple and fast to use, making it possible to manipulate the distribution of target variables to rebalance the dataset and increase the model's performance. Its ease of use and speed are a boon to data scientists managing a large-scale project.

Data Augmentation

Although the amount of data generated is increasing drastically, users are currently more concerned about how third parties collect, share and use their data. Ethical concerns take on greater salience, legal compliance requirements become more strict [10], and customers can even lose trust in companies that collect more of their data than needed [3].

In other words, although there is more data available, the data collection and preparation process can become arduous and expensive. Additionally, businesses might be facing situations where data projects must move forward with small datasets. Simulated datasets, if generated accurately, can provide meaningful insights [4] and help overcome problems such as shortage of data, whilst preserving customer privacy.

One of the main drawbacks of small datasets is that it becomes more difficult to avoid over-fitting and therefore fitting complex models such as Neural Networks becomes even more laborious. Models trained on smaller datasets have higher variance as they are highly affected by small perturbations such as outliers or noise.

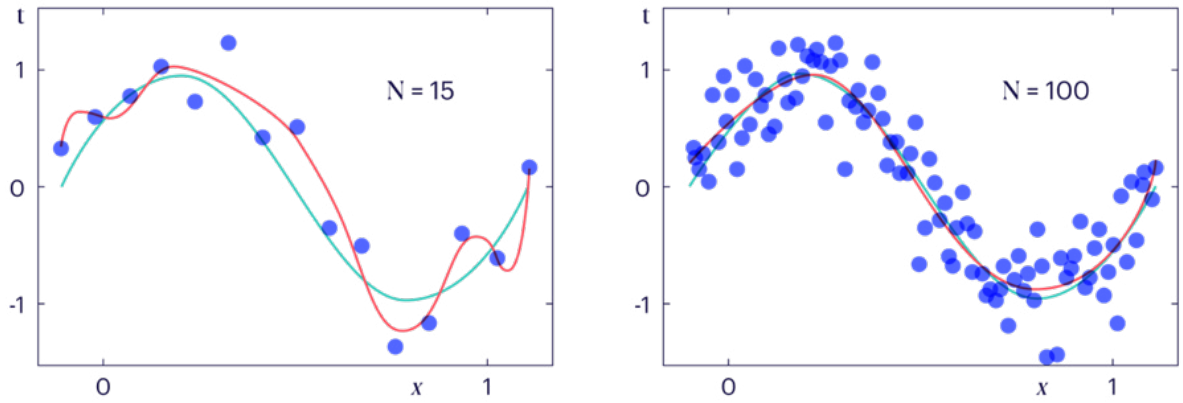


FIGURE 4.
N samples (blue circles) are drawn from a sine signal (green line) with additive noise. A ninth degree polynomial function (red line) is fit by minimising the sum of the squared errors for N=15 (left) and N=100 (right). Image from [2].

The figure above illustrates this issue. N points are sampled from a ground truth line with additive noise. Then, a ninth degree polynomial function is fit by minimizing the sum of the squared errors on these samples. For the N=15 case (left), the regression overfits and is not able to learn the ground truth as accurately as the N=100 case (right).

To explore this problem and demonstrate how it can be overcome with Synthesized data, we set up an experiment as follows:

1. Split the dataset into the training and test sets with ratio 4:1.
2. Resample the dataset:
    - To reduce its size (under 100%), a few rows will be randomly deleted from the training set.
    - To increase its size, we use Synthesized to learn the structure of the data, generate more samples, and append them to the training set.
3. Compute the evaluation metrics on the test set that remains unseen.

Figure 5 shows 10 Monte Carlo simulations of this procedure for the Absenteeism dataset, which contains only 740 rows and 21 columns. For the top two plots, each line corresponds to a different experiment, and the bottom two show the standard deviation at each step.

As can be observed on the top two plots, the average of both F1 and AUC is quite similar for all sampling proportions, so augmenting the dataset doesn't show an improvement on the average performance. Bottom two plots contain the standard deviation for all the simulations showing a clear downtrend on the stability of the metrics as we add Synthesized data to the training set. Especially in the AUC score (left), the experiment results converge to smaller variance (top) and the standard deviation is drastically reduced as the sample size increases (bottom).

The source of this instability is overfitting. Adding Synthesized data to the training set reduces variance, augmenting model's stability and reducing its sensitivity to outliers. In other words, adding Synthesized data to the training set can be thought of as another form of regularization.

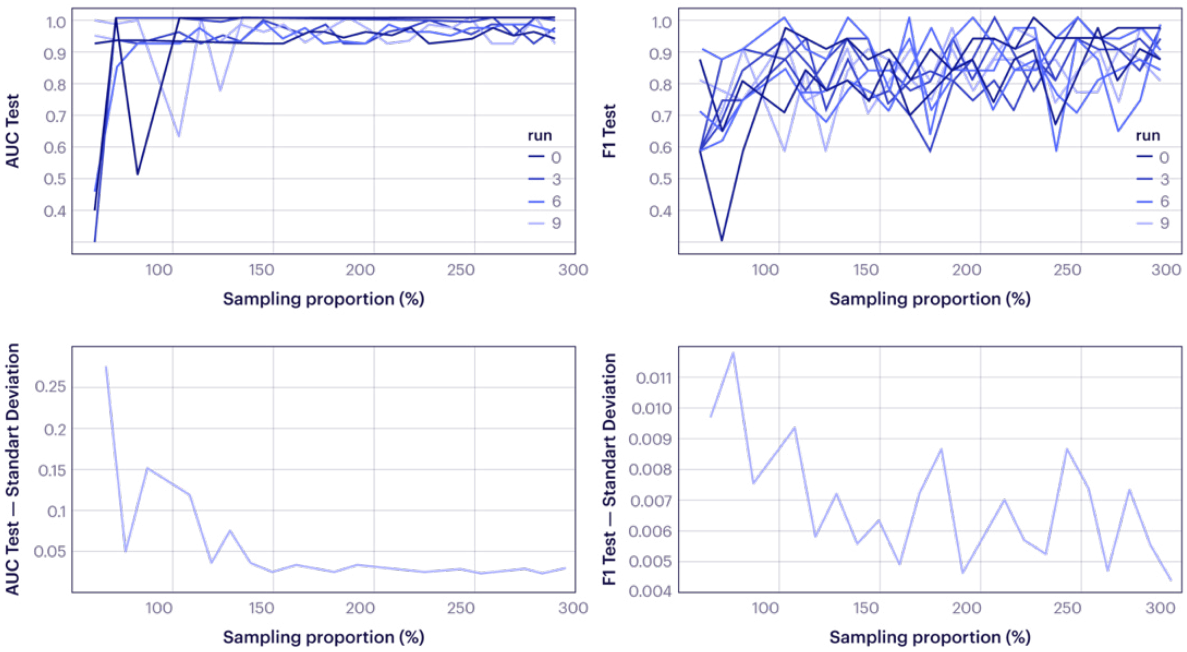


FIGURE 5 (PREVIOUS PAGE).
This graph shows the performance progression in terms of AUC (left) and F1 (right) for the *Absenteeism dataset* with different sample sizes. The top two plots show the results for 10 experiments with different random seeds, and the standard deviation of these experiments is displayed in the two bottom ones. The model used here is a Neural Network with layer sizes 64, 32 and 16. The X-axis represents the sampling proportion, where after 100% the data added is Synthesized.

In the case described above, this regularization factor added by Synthesized data helped to make model learning more stable, but in some other cases, it additionally shows an improvement on the performance metrics. The Japanese credit application dataset has been sampled from 70% to 300% of its original size to train a Gradient Boosting Machine, and the metrics are computed again in an unseen test set.
The results are displayed in Figure 6.

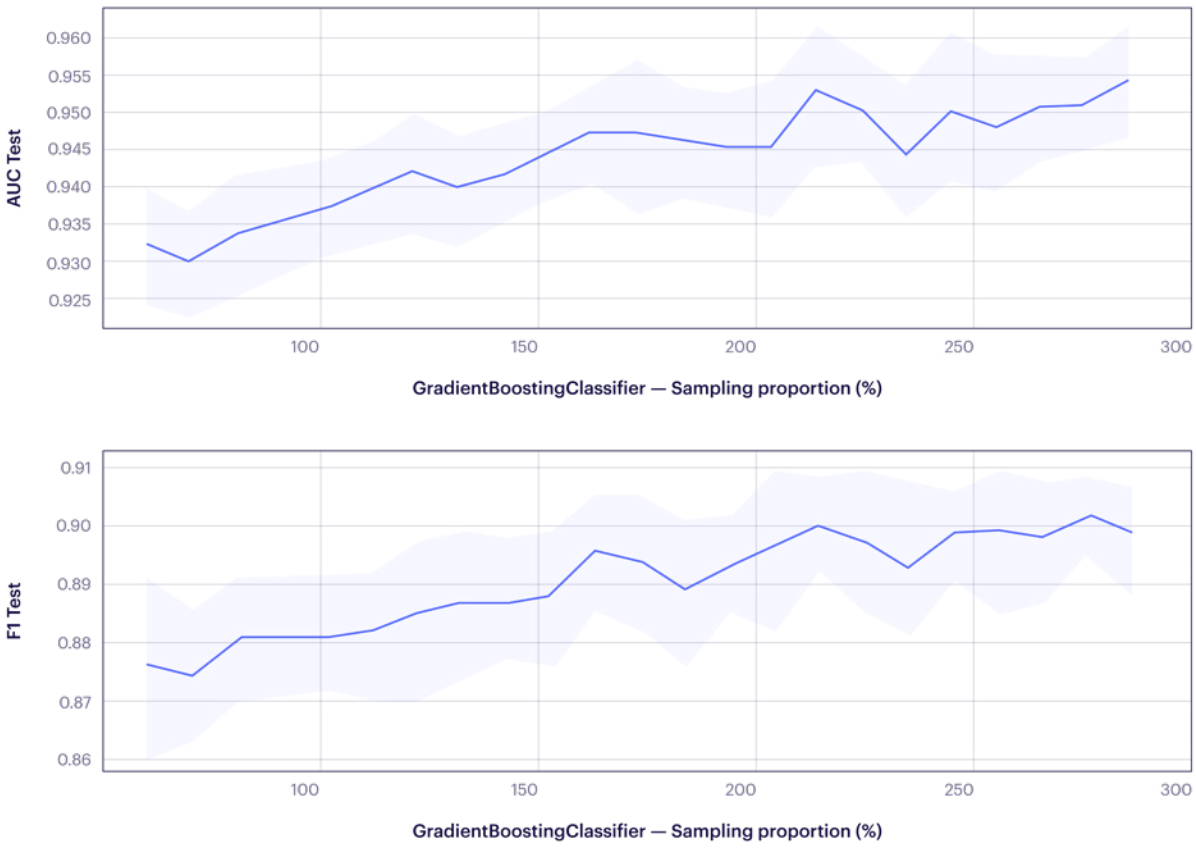


FIGURE 6.
The progression of the F1 (top) and AUC (bottom) metrics for a Gradient Boosting Classifier trained on the *Japanese credit application* dataset for different sample sizes from 20% of the original size to 500%.

For a credit dataset with ten thousand rows, the results still show an improvement on the resulting evaluation metrics. In this case, a neural network has been fitted again on the data resampled, and the outcomes are presented in Figure 7. On the top left, the AUC shows a clear increase as the sampling proportion augments, as can also be observed on the ROC curve on the right, where the original results are compared to three different samples. On the bottom left, the F1 score also shows a slight increase on the average performance.

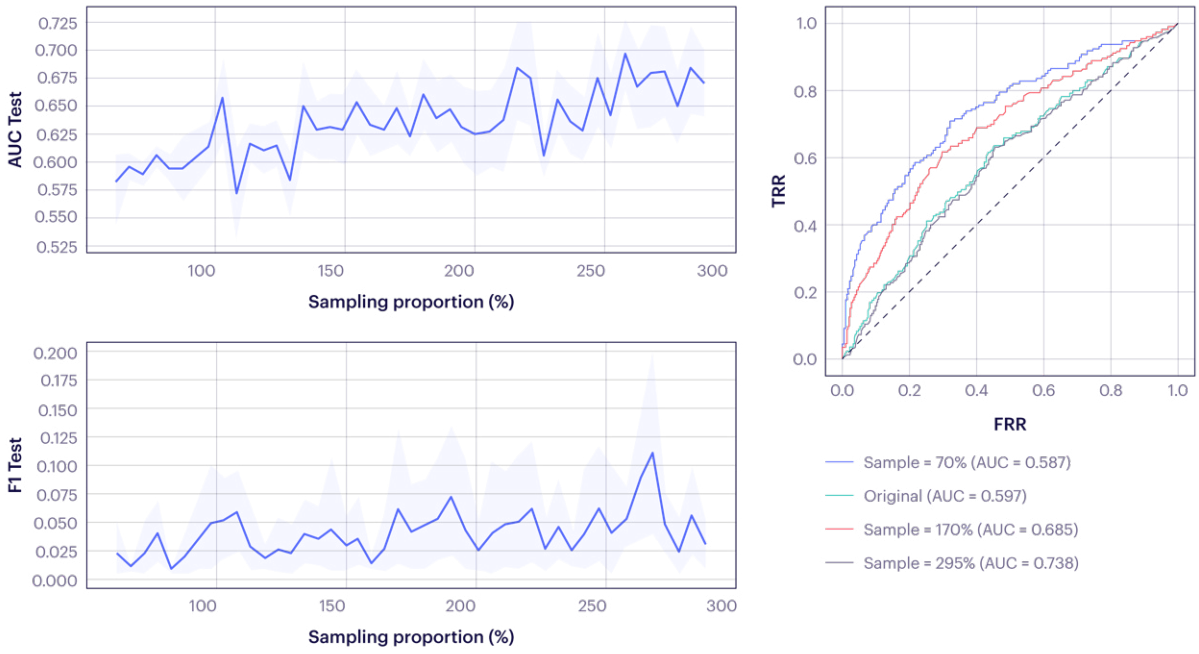


FIGURE 7.
A Neural Network trained on the credit dataset for different sample sizes. On the left, the AUC (top) and F1 (bottom) progressions, and on the right the ROC curve for three cases plus the original.

In sum, this section has demonstrated the benefits of using Synthesized for Data Augmentation, showing that it is another form of regularization. It ensures model stability for small datasets where one can easily overfit, but it can also improve model performance.

Model Stability Under Population Shifts

Customer behaviour predictive models, such as credit scoring or targeting, work under the assumption that past behaviour will be repeated in the future. But as the world changes, human behaviour can change, and an improperly validated model can lead to costly consequences [5] such as:

- Fast degradation of the model performance.
- Predictions that are overly sensitive to small input variations.
- Complete break-up under heavy population shifts (such as the current COVID-19 crisis).

With proper model validation, the effects of these shifts can be minimized. In this section, we compare the stability of two models (a Random Forest and a Neural Network) and analyse how they behave if the test population is altered. To do so, we use Synthesized to generate three different scenarios, while keeping the training set the same.

The results for both models trained and tested on the raw credit scoring dataset are the following:

	RANDOM FOREST	NEURAL NETWORK
F1	0.28	0.35
AUC	0.81	0.79

TABLE 2.
F1 and AUC scores for two models trained on the credit scoring dataset.

Looking at these metrics, one may think that the Neural Network is the best model to use, but a proper model validation should be performed before deploying this model. Here, Synthesized is used to generate two scenarios, one where the default rate is decreased to 5%, and another to 20%.

The ROC curves for these scenarios and both classifiers are shown in Figure 8. Random Forest shows much more stable behaviour than the Neural Network, as the performance of the latest is highly degraded by when the population changes. Similarly, the confusion matrices can be observed in Figure 9. The false negative error (usually the most expensive one) increases in both cases when the default rate goes up, but whilst the Random Forest only has 121 samples in this kind of error, the Neural Network has 432.

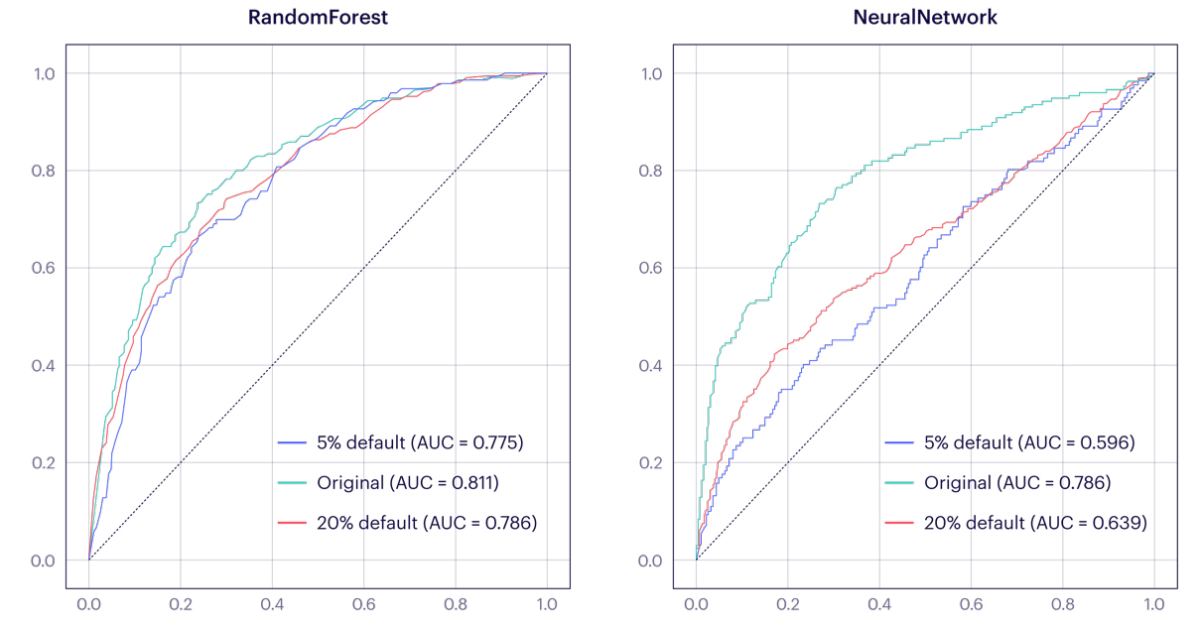


FIGURE 8.
Comparison of two models trained on the same credit dataset and tested on three different scenarios, (i) the original test set, (ii) a Synthesized dataset with 5% default rate, (iii) a Synthesized dataset with 20% default rate.

FIGURE 9 (PREVIOUS PAGE).
The confusion matrices for the predictions of a Random Forest (top) and a Neural Network (bottom) trained on the same credit dataset and tested on three different scenarios, (i) the original test set, (ii) a Synthesized dataset with 5% default rate, (iii) a Synthesized dataset with 20% default rate.

A data scientist has to properly validate a model before deploying it in production. In this section, we have shown how Synthesized can also be used to simulate hypothetical scenarios with the data manipulation toolbox, and ensure the model's stability under unexpected population shifts.

References

[1] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, pp.321-357.

[2] MLA. Bishop, Christopher M. Pattern Recognition and Machine Learning. New York: Springer, 2006.

[3] https://www.accenture.com/gb-en/insights/software-platforms/building-data-ai-ethics-committees

[4] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

[5] https://www.gigabitmagazine.com/big-data/dispelling-misconceptions-about-synthetic-data-sets

[6] https://www.forbes.com/sites/forbestechcouncil/2019/04/03/why-machine-learning-models-crash-and-burn-in-production

[7] https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

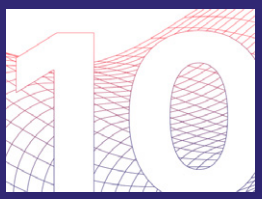
[8] http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303

[9] https://www.itproportal.com/features/why-digital-businesses-should-understand-their-data-better/

[10] https://financialit.net/blog/big-data/big-data-doesnt-need-be-big-or-daunting

SYNTHESIZED

YOU MAY ALSO BE INTERESTED IN:



Top 10 Synthetic Data Use Cases and Applications for 2022

Read Now →

AI and Data in Scotland: A Conversation with Gillian Docherty

Listen Now →

INTERESTED TO LEARN MORE?

If you would like a demo about our platform capabilities or would like to try it for free, please get in touch.

Contact our Machine Learning team for a personalised demo.

Request a Demo

Try the Synthesized SDK now at no cost.

Try Now

The **Synthesized** all-in-one DataOps Platform provides a comprehensive solution to the problem of collaboration and sharing sensitive data with high-quality fully-compliant data products.

team@synthesized.io

SYNTHESIZED