ANITA

Anonymous big data A project funded by FFG

New or refined model architectures

Deliverable D5.2

Author(s): Michael Platzer, Klaudius Kalcher

Reviewer(s): Peter Eigenschink

Document version: 0.2
Date: 07.06.2021

WU Wirtschaftsuniversität Wien Vienna University of Economics and Business · MOSTLY AI · George Labs · Statistik Austria Die Informationsmanager · FFG Promoting Innovation. · Federal Ministry Republic of Austria Climate Action, Environment, Energy, Mobility, Innovation and Technology

Programme **"ICT of the Future"** - the initiative of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK)

ANITA
Anonymous
big data

Disclaimer

This deliverable describes the work and findings of the AI-Based Privacy-Preserving Big Data Sharing for Market Research (Anonymous Big Data (ANITA)) project.

The authors of this document have made every effort to ensure that its content was accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this deliverable are responsible for any possible errors or omissions as well as for any results and actions that might occur as a result of using the content of this document.

Table of contents

1 Summary

An extensive list of model refinements has been concepted, developed and benchmarked as part of WP5, that further improved MOSTLY AI's existing architecture to serve the captured use cases and requirements:

- Improve accuracy of recurring high cardinality features, like Transaction Categories
- Improve accuracy of free text attributes, like Transaction Text
- Improve accuracy of geo-spatial information
- Improve training on larger datasets
- Improve rule adherence with conservative sampling strategies
- Explored handling of very wide tables with Regressor Attention
- Explored impact of embedding heuristic
- Explored impact of differential-private Stochastic Gradient Descent

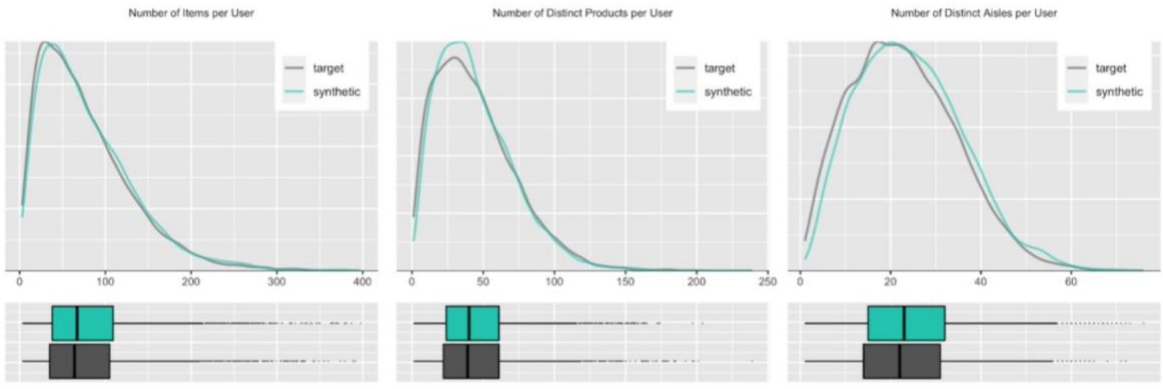
Please note, that key findings and corresponding research reports are project internal, and not contained within this document.

2 Researched Topics

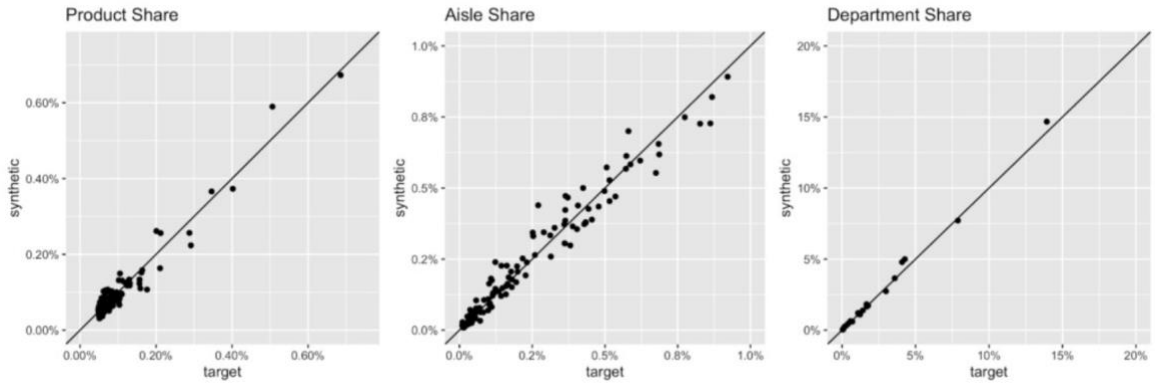
2.1 High Cardinality Features

WP4 introduced accuracy metrics that measure the coherence within sequential data (see D4.1). Empirical experiments show that it is of particular challenge for synthetic data solutions to remain coherent, if high cardinality features are present in the original data. This is for example the case for transaction categories in the finance industry, where each transaction can be assigned to any of over 100,000 categories. Or for the Instacart dataset (https://www.kaggle.com/c/instacart-market-basket-analysis), that contains purchases across over 50,000 distinct products. While univariate statistics are easy to capture, the challenge is to remain consistent across multiple events for a synthetic subject. I.e., if a synthetic customer has purchased Organic Avocados before, then that customer is likely to purchase Organic Avocados again. This tendency to re-purchase already purchased items needs to be learned for each product item independently by the model, and thus is not always reliably captured, particularly for less frequently purchased items.

As part of WP5 we've concepted, developed and benchmarked a novel approach that is capable of explicitly taking the information on already previously generated synthetic events into account, and then results in highly coherent transaction histories. This is achieved by explicitly taking the information into account (via a boolean flag) whether an item has already occurred or not, both for model training and for data generation. Detailed empirical results have been published at https://mostly.ai/2020/06/05/how-to-unlock-your-behavioral-data-assets-part-iii/



Side-by-side comparison of coherence metrics for high-cardinality attribute product for Instacart dataset.



Visualizations of high alignment in "market share" between original and synthetic data, despite the high cardinality in product items, for Instacart dataset.

2.2 Free Text

A key take-away from the captured use cases was that valuable statistical information is commonly contained not only in the structured fields, but also within the unstructured attributes, such as the text column of a financial transaction. These attributes can contain any possible combination of characters and are thus particularly challenging to learn and to retain in the context of all other attributes. A subsequent challenge is then the measurement of accuracy of these unstructured attributes.

As part of WP5 we've concepted, developed and benchmarked a novel approach to combine the synthetization of structured mixed-type sequential data, with unstructured attributes. We take advantage of existing text tokenization libraries (Google's SentencePiece, HuggingFace's Tokenizers, etc.), that can automatically adapt to the domain of the data, and then transform the data into a particular data structure for the model training. Once, synthetization is complete, the transformation is being reversed, resulting in highly representative, and highly realistic free text attributes.

These are some non-cherry-picked examples of generated synthetic tweets, based on a model trained on 1.6 million actual tweets (https://www.kaggle.com/kazanova/sentiment140):

- *Watching the Procentric channel with my baby*
- *Good morning everyone*
- *@Paincat well, I am working tonight so I have to leave today early: carrriiiiiiiiiieeeeeeeeeeeeeeeeeeee*
- *@Xsid hey! thanks for following*
- *@jeffsamney anytime Happy Birthday! you might get a comment*
- *What a busy day. Some things all works out. And no Saga!*

Quality assessment is done by either calculating basic word statistics and co-occurrences:

	1x@	2x@	#	http:	work	tomorrow	work|tomorrow	.com|http	!	!|lol	)|(
original	0.3834	0.0135	0.0249	0.0462	0.0573	0.0241	0.1551	0.6371	0.2907	0.3639	0.7373
synthetic	0.4012	0.0193	0.0291	0.0440	0.0586	0.0224	0.1167	0.6091	0.3074	0.3936	0.7160

Or by looking at class distributions, of a downstream classifier applied to the generated data. In this case we used a sentiment classifier.

sentiment (3-class)	MAE	negative	neutral	positive
actual	-	32%	31%	37%
synthetic	**3.6%**	28%	36%	35%

2.3 Scale Training to Larger Data Volumes

Being able to process more training data enables deeper levels of statistical patterns to be reliably captured. Thus, we investigated various ways to feed more data into the training process:

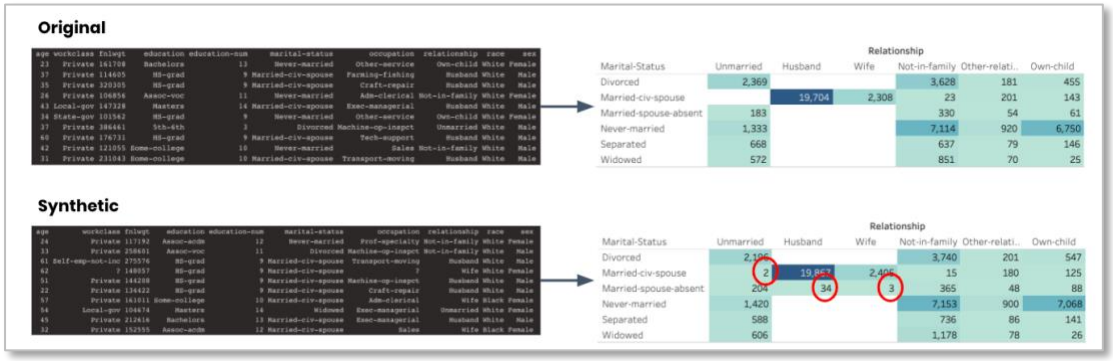
- We conducted experiments on large batch_size training
  - for single table
  - for sequential data
- We concepted and prototyped a smarter early stopping to drastically reduce total training time
- We prototyped Multi-GPU support for model training

Particularly for single-table setups we were able to significantly speed up training time by opting for larger batch sizes without losing on accuracy.

2.4 Rule adherence

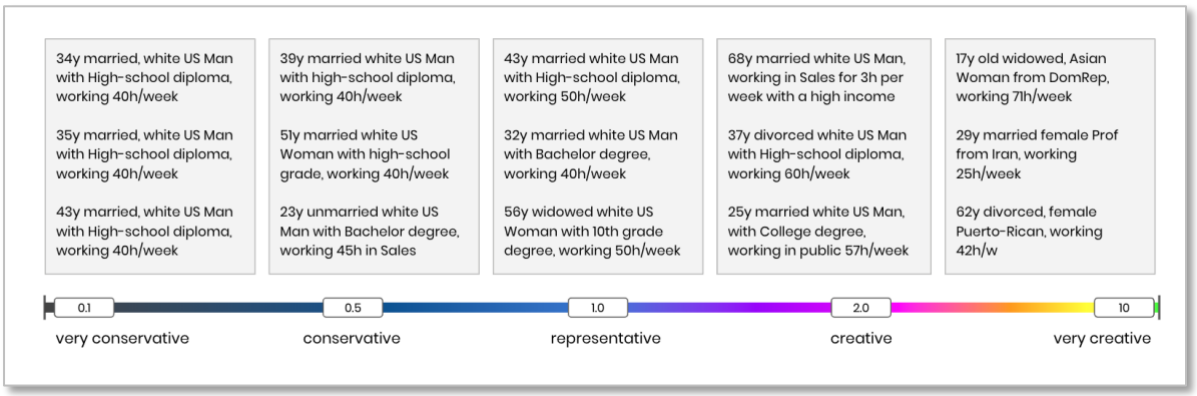
Use cases around testing & development require less emphasis on statistical representativeness, but more on the plausibility of generated synthetic records. While it is of value to generate new, yet unseen value combinations, it is counterproductive to have synthetic test data generated that violates existing business rules.

We first analyzed existing datasets and checked for the presence of business rules. For example, the UCI adult dataset (https://archive.ics.uci.edu/ml/datasets/adult) exhibits clear rules. Records that have their marital.status set to divorced, shall not be flagged as Wife or Husband with respect to their relationship.



As part of WP5 we've concepted, and protoyped multiple strategies to ensure that existing rules are automatically detected and adhered to. In addition, we've explored conservative sampling techniques (Holtzman et al. 2020) for the generation process itself, that allow to significantly reduce the chance of violated business rules. In fact, this would allow users to gradually trade-off the statistical representativeness of synthetic data for a higher chance to have only rule adhering records generated.

The following figure displays such a trade-off demonstrated on top of synthetic US census data. The samples on the left hand side represent samples biased towards the most likely values, and the right hand side allows for less likely value combinations to occur. If rule adherence is an explicit goal, then shifting the sampling temperature to lower values can achieve that.



Samples of synthetic US census records, generated at varying degree of temperature.

2.5 Support for very wide tables

For analytical use cases it is common to have data denormalized across multiple tables, ending up with very wide data tables. The challenge is then to synthesize these wide tables of up to 1,000 columns, while still being able to retain the statistical patterns among all of these. Even when looking at only bivariate relationships, the number of possible combinations grows by the square (e.g. 500,000 = 1,000 x 1,000 / 2). It is important to systematically check the accuracy across multi-variate relations.

We researched novel approaches leveraging Regressor Attentions, i.e., an attention layer on top of the final regressor layers of the preceding columns, to support the model learning relationships across many variables. The results were analyzed in terms of accuracy, privacy and computation costs.

BackBlaze Harddrive Data - 2020-01-01

Experimental Set Up

- Default Values
- number_of_regressor_units [20,50]

Results

Epoch Duration

branch_type	amax	amin	mean	median	std
Attention (20)	41451.7	378.658	1446.26	389.457	6574.58
Attention (50)	38517.7	389.469	978.602	401.464	4691.84
Current (20)	1909.57	72.3914	109.541	78.4646	242.698

Validation Loss

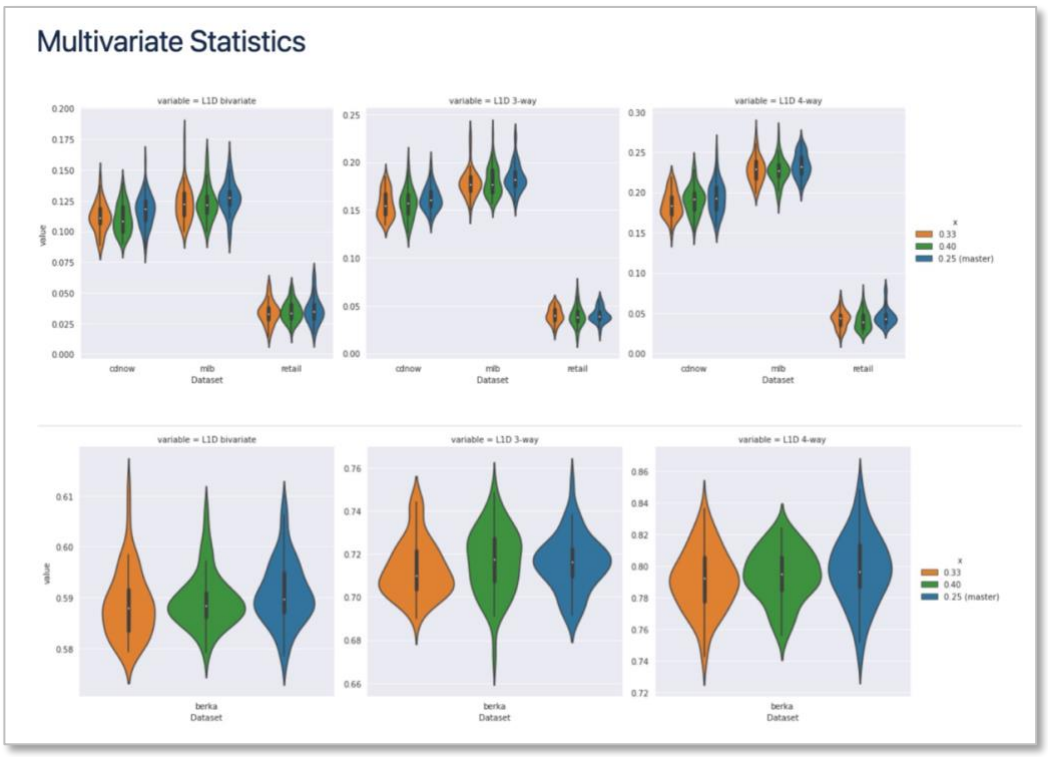
branch_type	amax	amin	mean	median	std
Attention (20)	171.137	119.47	125.889	121.99	10.2058
Attention (50)	167.997	114.381	119.29	116.47	8.4596
Current (20)	154.238	114.443	118.554	116.27	6.63492

Snippet from internal research report, comparing numbers with and without regressor attention.

2.6 Embedding heuristic

High-cardinality attributes are typically passed through an embedding layer, that compresses one-hot encoded information into a task-efficient lower-dimensional representation. While there exist rules of thumbs within the deep learning literature, it is unclear how small or large that embedding layer should ideally be sized in practice.

We therefore leveraged the Virtual Data Lab to explore various rules, that set the size of the embedding layer in dependency of given data-related statistics.



Snippet from internal research report, comparing numbers with varying sizes of embedding layers. Lower L1D scores are better.

2.7 Differentially Private Stochastic Gradient Descent

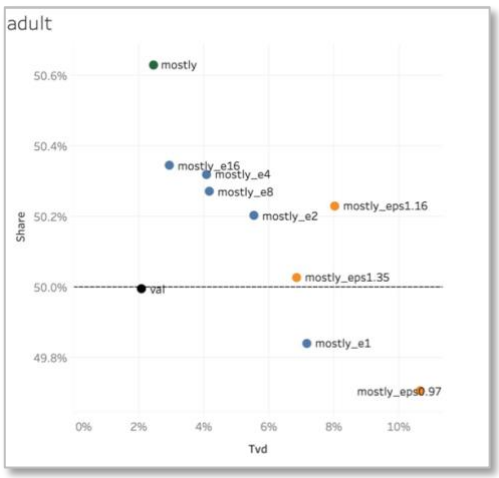
Differential Privacy (Dwork & Naor, 2006) is a mathematical concept providing an upper limit Epsilon for a given algorithm for how much any individual, present or not in a dataset, may alter its results.

It thus represents one out of several alternative privacy concepts.

- Differential Privacy offers mathematical guarantee with clear definition and elegant mathematical properties
- However, it doesn't necessarily relate to empirically relevant privacy criteria
- No common agreement on the level of acceptable Epsilon
- Differential Privacy guarantees are a property of the algorithm and cannot be empirically validated based on the results of the algorithm alone
- Thus, no regulation has currently adopted differential privacy

Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al. 2016) is one approach that allows to make the Stochatic Gradient Descent computation within the training of deep neural networks differentially private. It aims to preserve privacy by gradient norm clipping and by adding noise to it.

We've implemented a DP-SGD variant of and explored various settings for the corresponding hyper parameters (microbatch_size, noise_multiplier, clipnorm, vector-optimized). Experiments on top of Virtual Data Lab show that DP-SGD comes with a significant computational overhead, while also negatively impacting the overall accuracy. Plus, the developed empirical privacy tests (https://arxiv.org/abs/2104.00635) do not seem to benefit from imposing this theoretical mathematical guarantee.



Side-by-side comparison of empirical fidelity (x-axis) and empirical privacy (y-axis) for US Census dataset across a range of synthetic data sets.

The figure below displays the empirical fidelity and empirical privacy across multiple synthetization runs for the US census dataset, that is part of the Virtual Data Lab (D4.1.). The blue dots represent synthetization runs that are stopped already after a handful of epochs (mostly_e**X** = stopped after X-th epoch). The orange dots represent synthetization runs with DP-SGD, and their corresponding epsilon values. It can be seen, that despite significantly higher computational efforts, these differentially private runs do not achieve improved scores when compared to a run, that is simply stopped after its first training epoch.