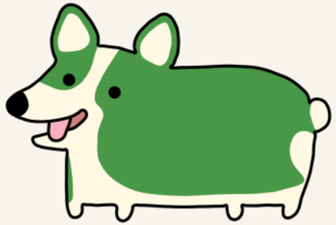
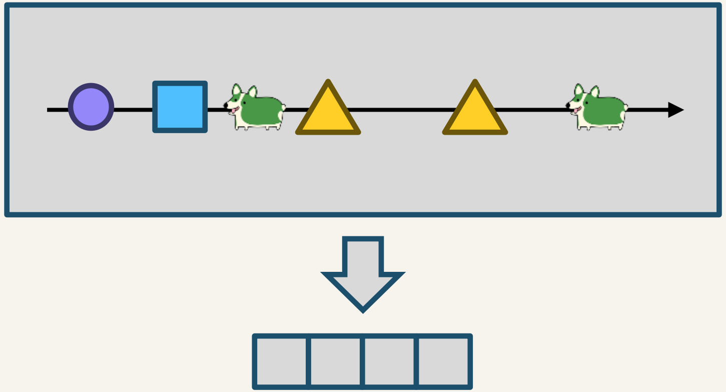
Universal representations for event sequences: financial transactional data and beyond



Alexey Zaytsev

Assistant professor
LARSS laboratory,
Skoltech



Skoltech

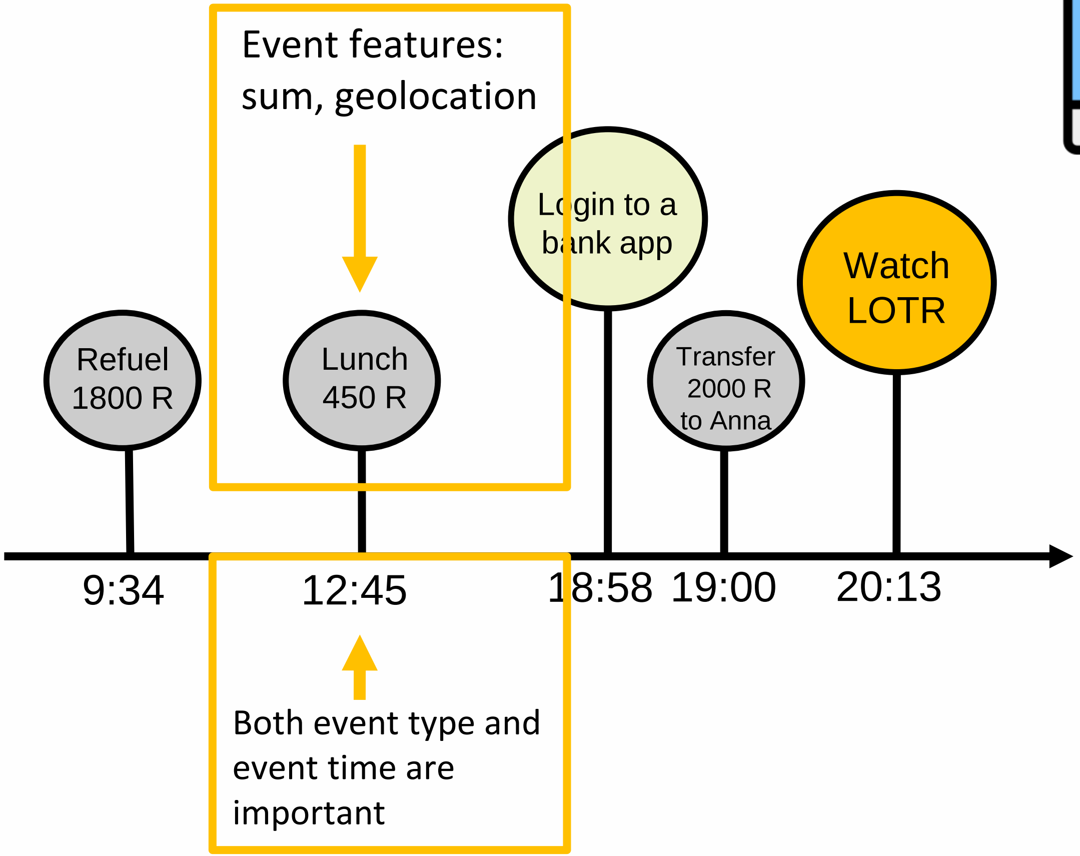
HSE
October 2024

Event sequences: intro

Skoltech

Event sequences data

Alex, 35, man
works in academia



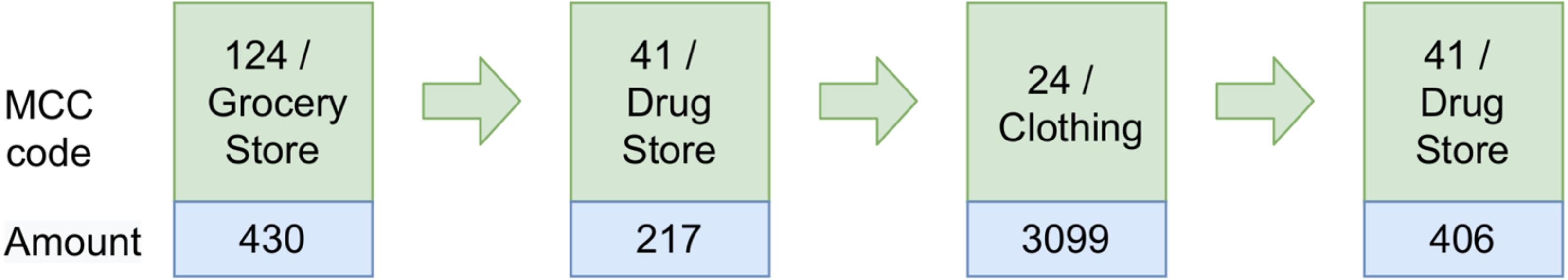
Discrete sequential transactional data

Transaction records data sequence includes:

- MCC (Merchant Category Codes)
- Purchase amount
- Time values
- Transaction location
- …

Data characteristics:

- Heterogeneous features
- Non-regularity of observations
- Varying lengths of sequences



Skoltech

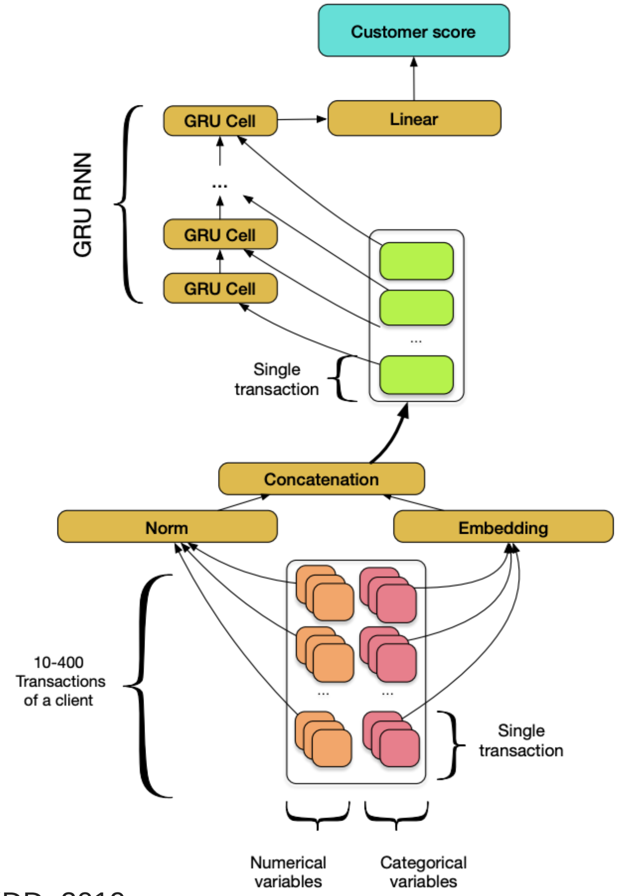
One way: Supervised approach

Library pytorch-lifestream

Recurrent (or Transformer) Neural Network with self-supervised contrastive learning

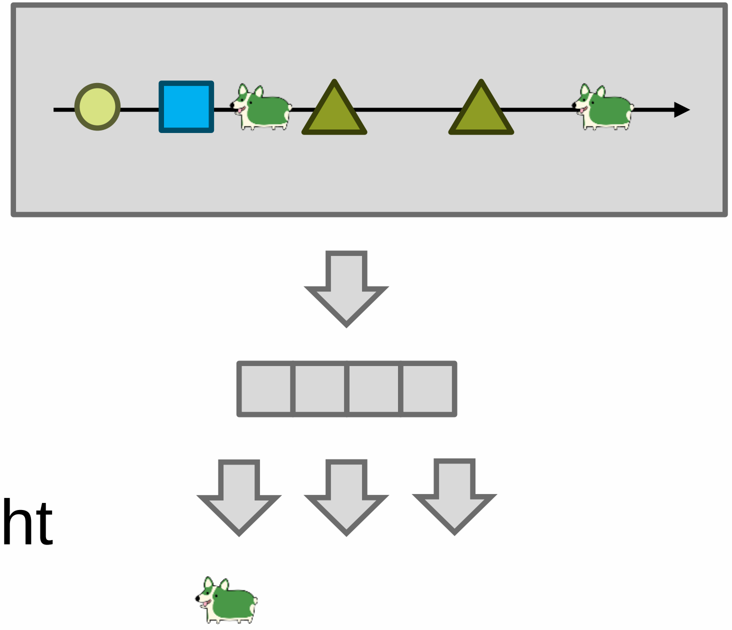
	ROC AUC	N Features
Logistic regression	0.78	~ 400
LGBM	0.81	~ 7000
E.T.-RNN	0.83	12

- Requires labeled data!



D. Babaev et al. E.T.-RNN: Applying Deep Learning to Credit Loan Applications. KDD. 2019

Our way: self-supervised approach



Encoder

Trained without labels

Lightweight heads

Trained using small number of labels

D. Babaev et al. E.T.-RNN: Applying Deep Learning to Credit Loan Applications. KDD. 2019

Types of self-supervised learning

Skoltech

Two main ideas of self-supervised learning: generative and contrastive

Self-supervised representation learning

Contrastive

Similar objects should have similar representation, different objects should have different representations

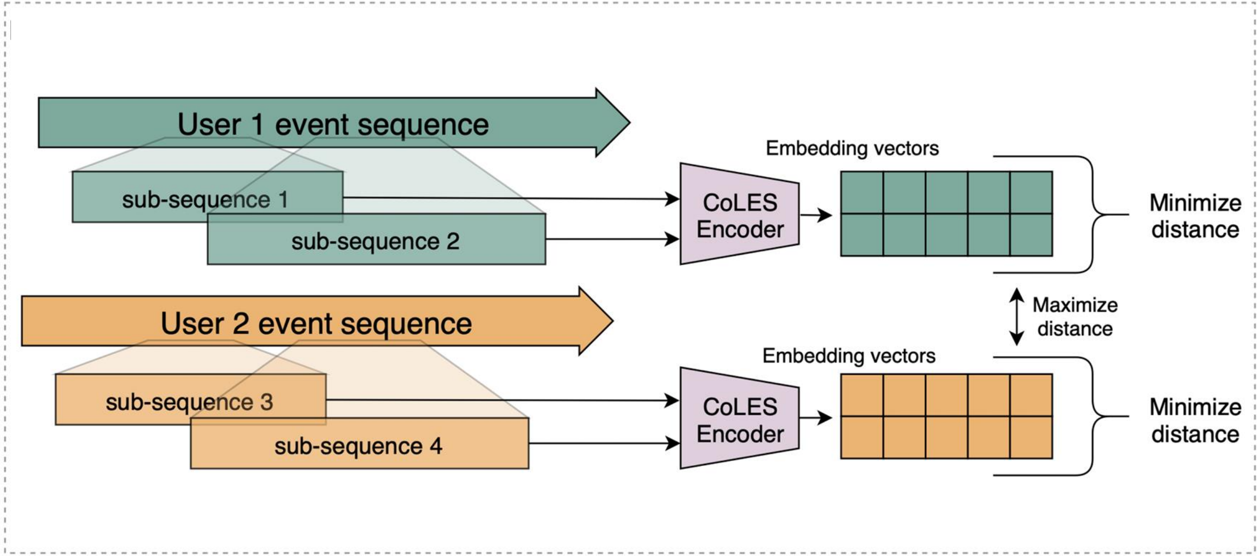
Generative

“Autoregression”: predict future using representations

Skoltech

Liu, Xiao, et al. "Self-supervised learning: Generative or contrastive." *IEEE Transactions on Knowledge and Data Engineering* (2021).

CoLES contrastive learning



Skoltech

Babaev, Dmitrii, et al. "CoLES: Contrastive learning for event sequences with self-supervision." *SIGMOD.* 2022.

Contrastive learning for sequential data

Weak augmentation: jitter-and-scale strategy
Strong augmentation: permutation-and-jitter strategy

$$\mathcal{L} = \lambda_1 \cdot \boxed{(\mathcal{L}_{TC}^s + \mathcal{L}_{TC}^w)} + \lambda_2 \cdot \boxed{\mathcal{L}_{CC}}$$

Predict future representation from the current context

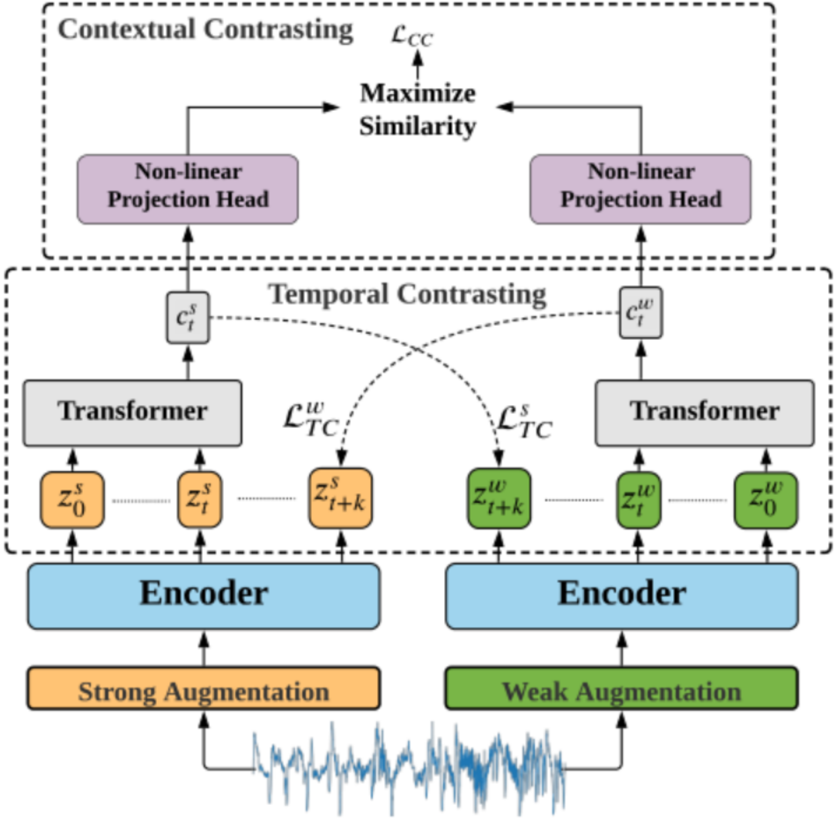
Compare contexts

Context:

$$\mathbf{c}_t = f_{AR}(Z_{\leq t}),$$

Representation:

$$\mathbf{z}_t = f_{enc}(\mathbf{x}_t)$$



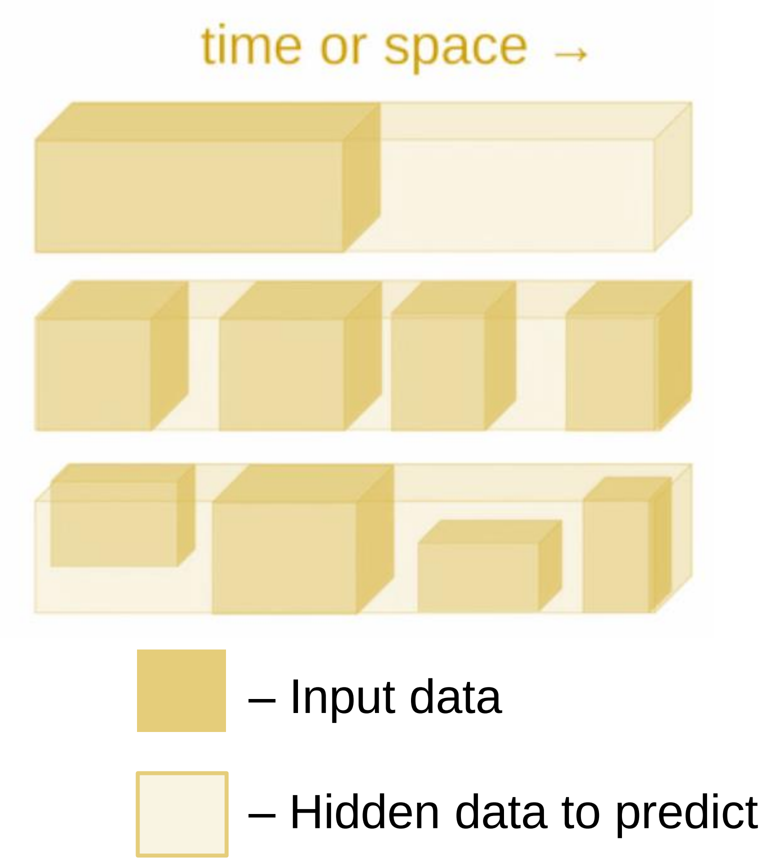
Skoltech

Eldele, Emadeldeen, et al. Time-series representation learning via temporal and contextual contrasting. *IJCAI. 2021*.

Generative models: masking

Training steps:

1. Hide some part of the data
2. Try to recover it via representation learning



A. Predict the future from the past

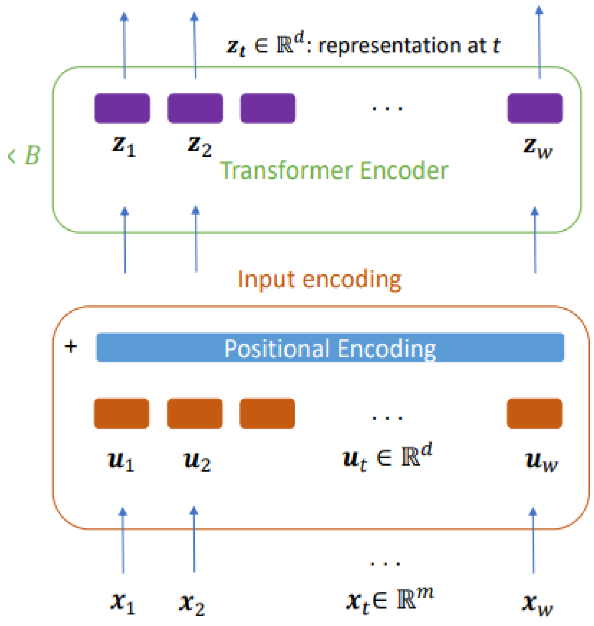
B. Predict the invisible from the visible

C. Predict occluded, masked or corrupted part

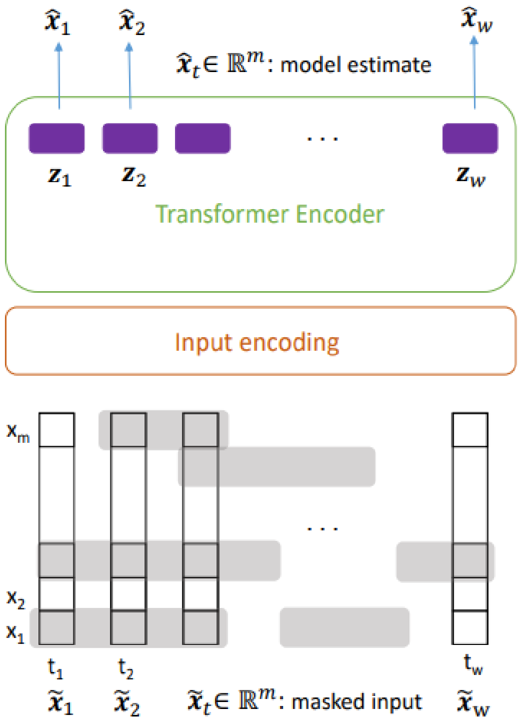
Skoltech

Yoshua Bengio, Yann LeCun Reflections (Self-Supervised Learning). ICLR 2020

Time-series unsupervised representations



Time series encoding via Transformers



Masking for model training

Zerveas, George, et al. A transformer-based framework for multivariate time series representation learning. KDD. 2021.

Skoltech

Desired properties of embeddings

Skoltech

Properties of event sequence embeddings

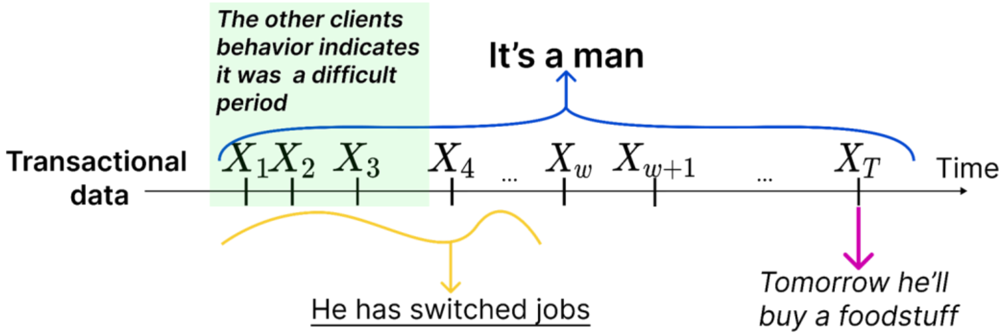
Goal: to obtain a good encoder for transactional data

Three main properties of local embedding for transactional data:

1. **Global property** - describe a client in general;

2. Local property – describe a client's state at a particular moment in time;

3. Dynamic property - the embeddings should change with time, reflecting the changes in the client's behavior.



Three main properties of embedding and corresponding business information: global (blue), local (pink) and dynamic (orange)

Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

Skoltech

Global and local quality of the models

- *Global validation* – solve a downstream task via a boosting model, get ROC AUC;

- *Local validation* – two approaches:
    a. predict the next event type (MCC) via MLP, get ROC AUC instead of likelihood**;**
    b. predict a local downstream target (churn/default state at the moment) via MLP, estimate ROC AUC**.**

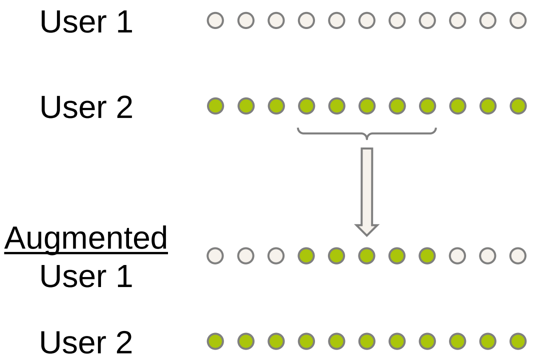
Skoltech

Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

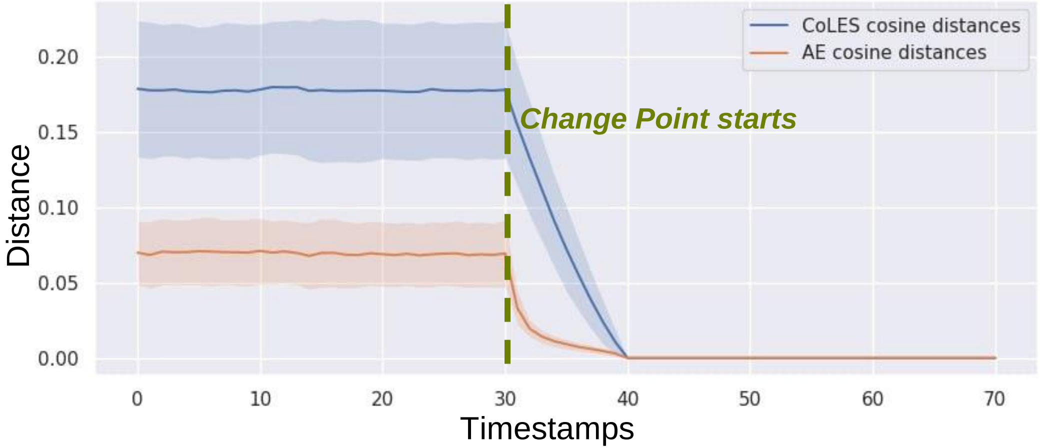
CoLES (contrastive) vs AE (generative): reaction to change

We also evaluate the models' **ability to detect user behavior change.** See an artificial change.

Experiment: "A poor man won a lottery".



Augmentation procedure. User 1 transactions were replaced with User 2 transactions. We compare User 2 to the augmented User 1.



Cosine distance between embeddings obtained from raw users and augmented ones. Snapshot near the Change Point

We expect embedding during the "augmented" area will be close to each other and far during other timestamps.

Global properties of models

Ranks for a local problem

	Age	Churn	Default	HSBC	Mean
AR§	1	1	1	1	1.00
MLM§	3	1	2	1	1.75
CoLES ext.†	3	2	2	3	2.50
AE§	2	3	4	2	2.75
CoLES†	3	4	3	3	3.25
Best baseline	4	4	5	3	4.00
TS2Vec†	6	4	5	4	4.75
A-NHP‡	5	5	6	4	5.00
NHP‡	5	5	6	4	5.00
COTIC‡	6	6	7	5	6.00

Ranks for a global problem

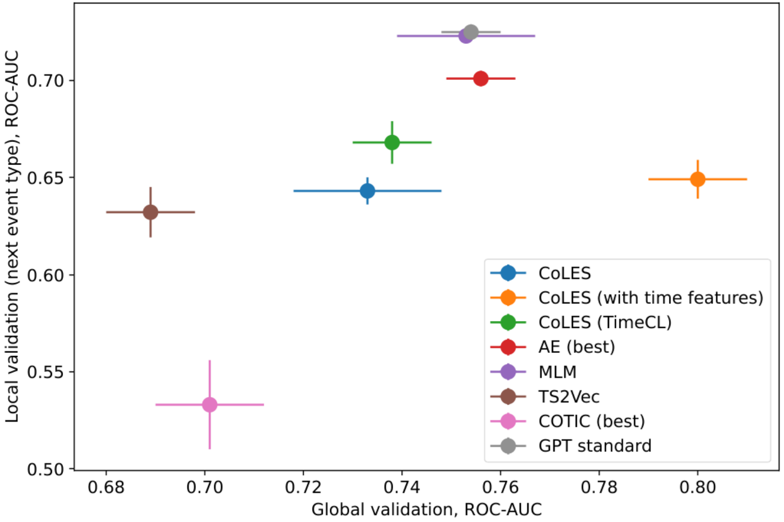
	Age	Churn	Default	HSBC	Mean
CoLES ext.†	1	1	1	1	1.00
CoLES†	1	3	1	1	1.50
MLM§	2	2	2	2	2.00
Best baseline	1	4	2	2	2.25
AR§	3	2	1	3	2.25
AE§	4	2	2	2	2.50
NHP‡	5	2	2	3	3.00
COTIC‡	6	3	1	4	3.50
TS2Vec†	2	5	2	5	3.50
A-NHP‡	5	3	3	4	3.75

Models are colour-coded: blue for generative,
green for contrastive and fuchsia for TPP.

Skoltech

Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

Comparison of local and global properties of models



Main conclusions:

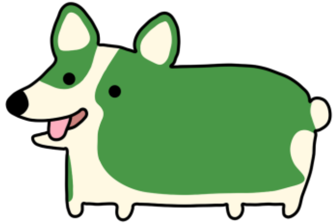
- GPT is better in local task.
- CoLES with time features is a clear leader in global validation.

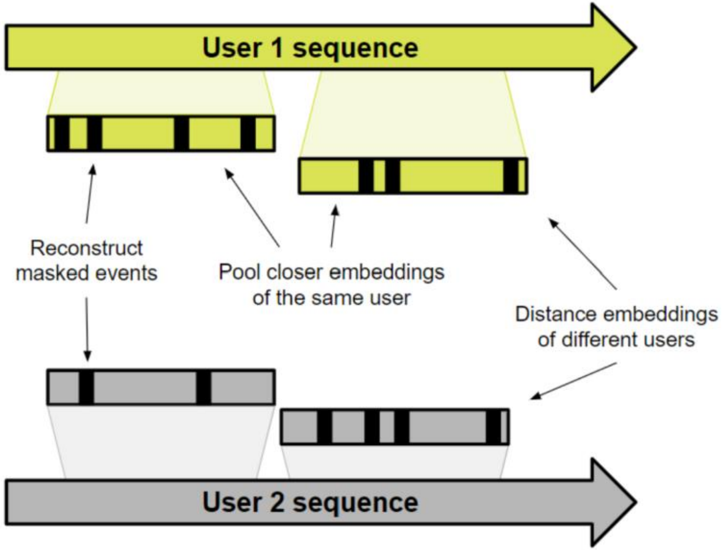
Bazarova, Alexandra, et al. "Universal representations for financial transactional data: embracing local, global, and external contexts." *arXiv preprint arXiv:2404.02047* (2024).

Combining contrastive learning and autoregression

Skoltech

Combining local and global properties





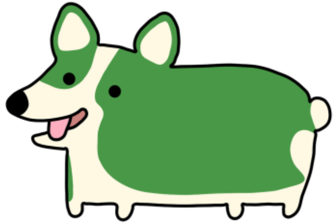
1. <u>Generative</u> reconstruction embeddings of masked events

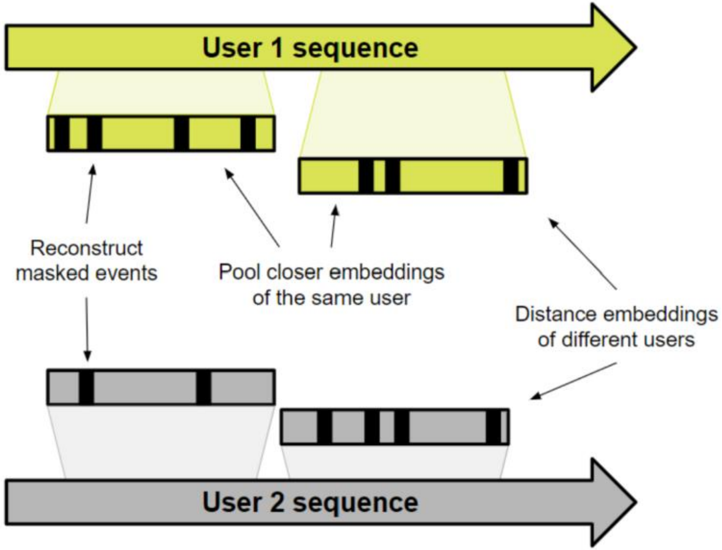
2. <u>Contrastive</u> comparison of embeddings from different users

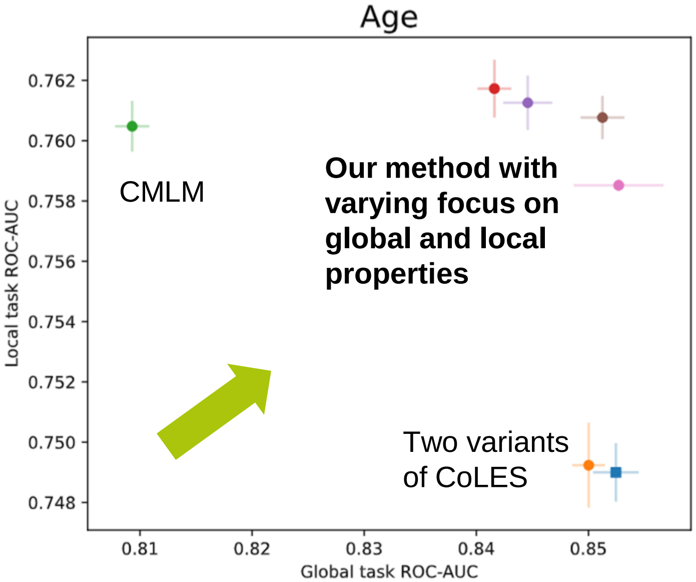
We simultaneously reconstruct embeddings with our CMLM and contrast in CoLES style

Yugay, Aleksandr, and Alexey Zaytsev. Uniting contrastive and generative learning for event sequences models. AIST. 2024.

Combining local and global properties



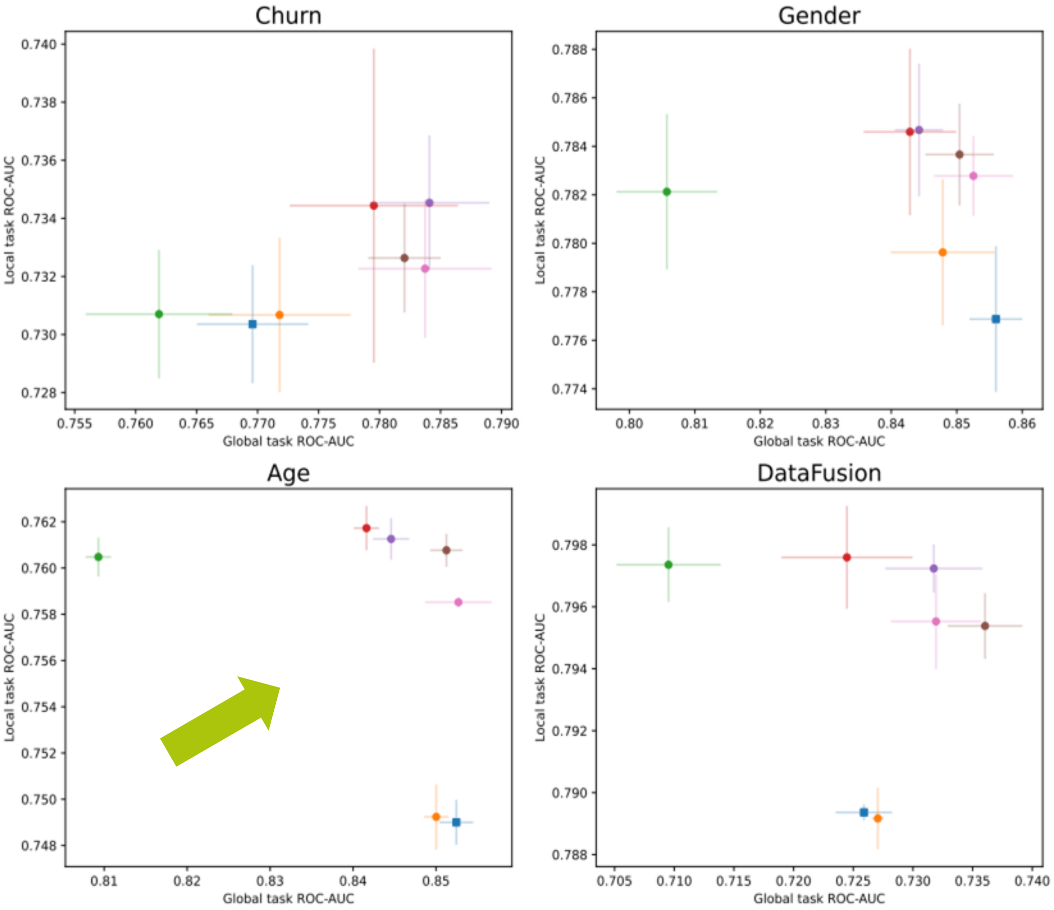


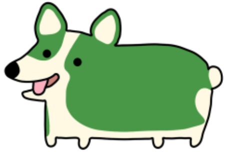


We simultaneously reconstruct embeddings with our CMLM and contrast in CoLES style

Skoltech

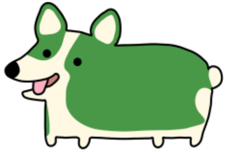
Results

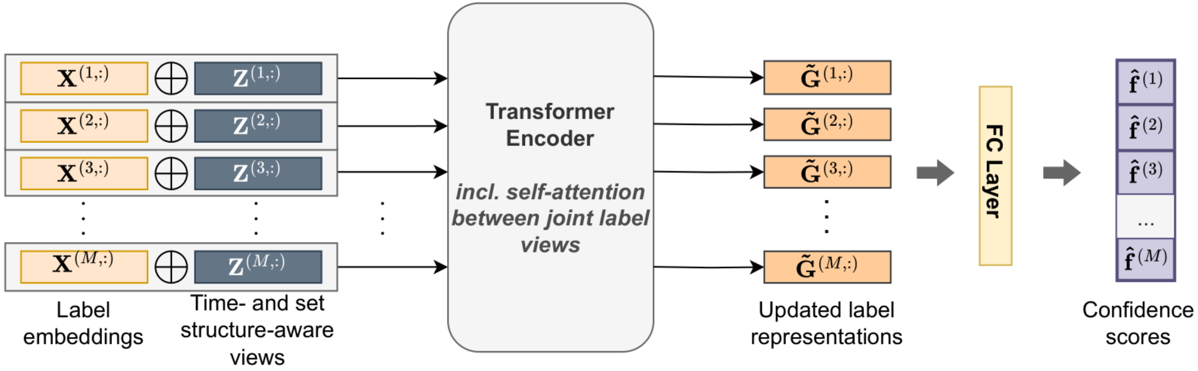




Method	*Global Task ROC-AUC*	*Local Task ROC-AUC*
Churn		
CoLES	0.770±0.007	0.730±0.003
CoLES (masking)	0.772±0.007	0.731±0.003
CMLM	0.762±0.010	0.731±0.004
CMLM+CoLES ($\lambda = 0.1$)	0.780±0.008	0.734±0.006
CMLM+CoLES ($\lambda = 0.05$)	**0.784±0.008**	**0.735±0.004**
CMLM+CoLES ($\lambda = 0.01$)	0.782±0.005	0.733±0.003
CMLM+CoLES ($\lambda = 0.005$)	**0.784±0.009**	0.732±0.004
Gender		
CoLES	**0.856±0.005**	0.777±0.004
CoLES (masking)	0.848±0.009	0.780±0.003
CMLM	0.806±0.009	0.782±0.004
CMLM+CoLES ($\lambda = 0.1$)	0.843±0.007	**0.785±0.004**
CMLM+CoLES ($\lambda = 0.05$)	0.844±0.004	**0.785±0.003**
CMLM+CoLES ($\lambda = 0.01$)	0.850±0.005	0.784±0.002
CMLM+CoLES ($\lambda = 0.005$)	0.853±0.008	0.783±0.002
Age		
CoLES	0.852±0.002	0.749±0.001
CoLES (masking)	0.850±0.001	0.749±0.001
CMLM	0.809±0.002	0.760±0.001
CMLM+CoLES ($\lambda = 0.1$)	0.842±0.002	**0.762±0.001**
CMLM+CoLES ($\lambda = 0.05$)	0.845±0.002	0.761±0.001
CMLM+CoLES ($\lambda = 0.01$)	0.851±0.002	0.761±0.001
CMLM+CoLES ($\lambda = 0.005$)	**0.853±0.005**	0.759±0.000
DataFusion		
CoLES	0.726±0.003	0.789±0.000
CoLES (masking)	0.727±0.001	0.789±0.001
CMLM	0.710±0.005	0.797±0.001
CMLM+CoLES ($\lambda = 0.1$)	0.724±0.005	**0.798±0.001**
CMLM+CoLES ($\lambda = 0.05$)	0.732±0.005	0.797±0.001
CMLM+CoLES ($\lambda = 0.01$)	**0.736±0.003**	0.795±0.001
CMLM+CoLES ($\lambda = 0.005$)	0.734±0.005	0.795±0.001

You were looking at a wrong self-attention?





We compute self-attention over event types and get prediction of next event type, imposing simple aggregation of temporal encodings.

Our LaNET model is now SOTA for the next basked prediction

Paper

GitHub

Kovtun, Elizaveta, et al. Label attention network for sequential multi-label classification: you were looking at a wrong self-attention. ECAI. 2024.

Few final words

Skoltech

Conclusion

- Typical SSL approaches focus on different aspects of embedding properties, also demonstrating generative capabilities

- We propose an SSL hybrid approach CMLM+CoLES that achieve notable improvements in both local and global properties of learned representations.

- Generative models for event sequences data are on their way!

Thanks my lab for help with these slides and you for your attention!

Alexandra Bazarova
Maria Kovaleva
Ilya Kuleshov
Evgenia Romanenkova
Alexander Stepikin
Alexandr Yugay
Elizaveta Kovtun
Galina Boeva
Andrey Shulga
Alexey Zaytsev

Skoltech

Thanks for your attention!

Backslides

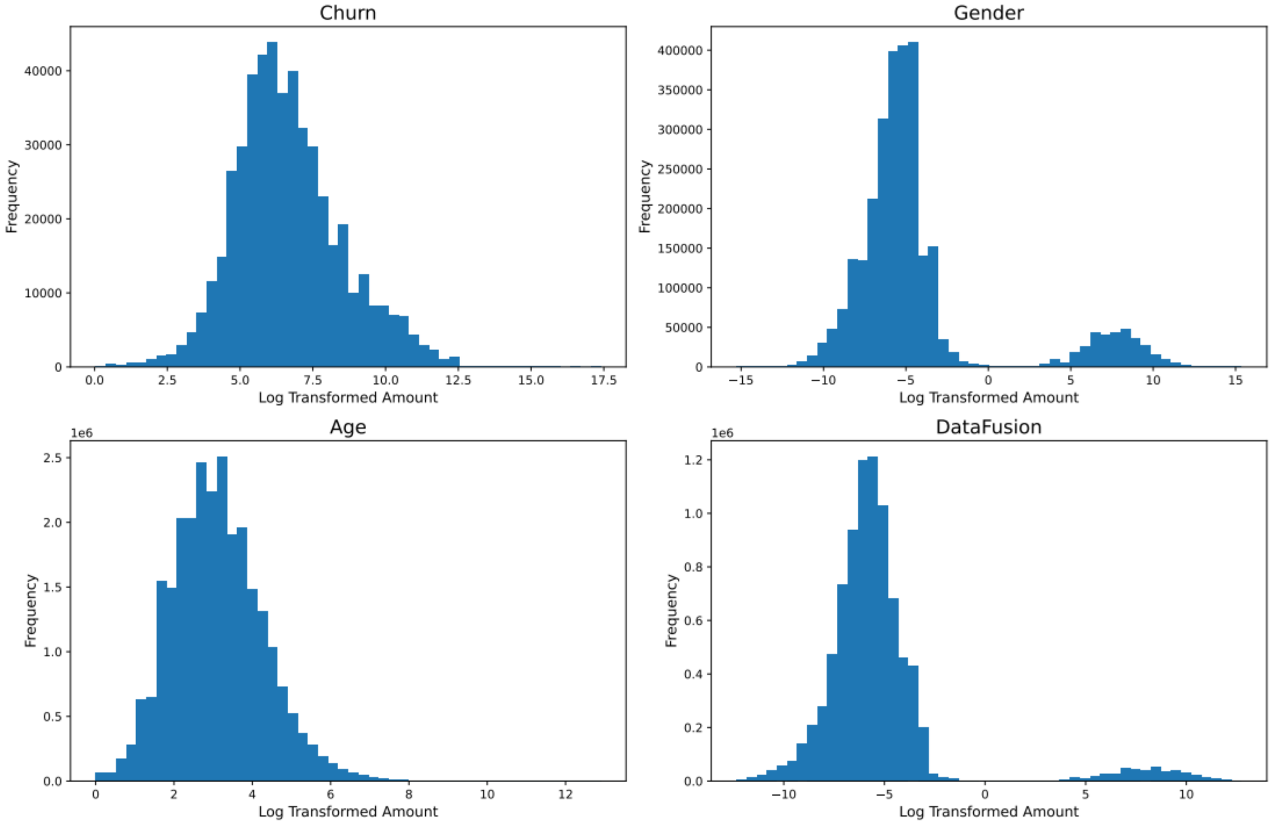
Experiment design

1. Pretrain models in a self-supervised regime
2. Use the obtained encoder as feature extractor
3. Train another model in a supervised regime on extracted features to solve downstream tasks:
   - Sequence classification
   - Next event type prediction

	Churn	Gender	Age	DataFusion
Num Transactions	490K	2.9M	26M	8.7M
Num Sequences	5K	7.4K	30K	64K
Mean Sequence Length	98.1	388.2	881.7	136.5
Std. Sequence Length	78.1	309.4	124.8	148.9
Num Unique MCC	344	184	202	323

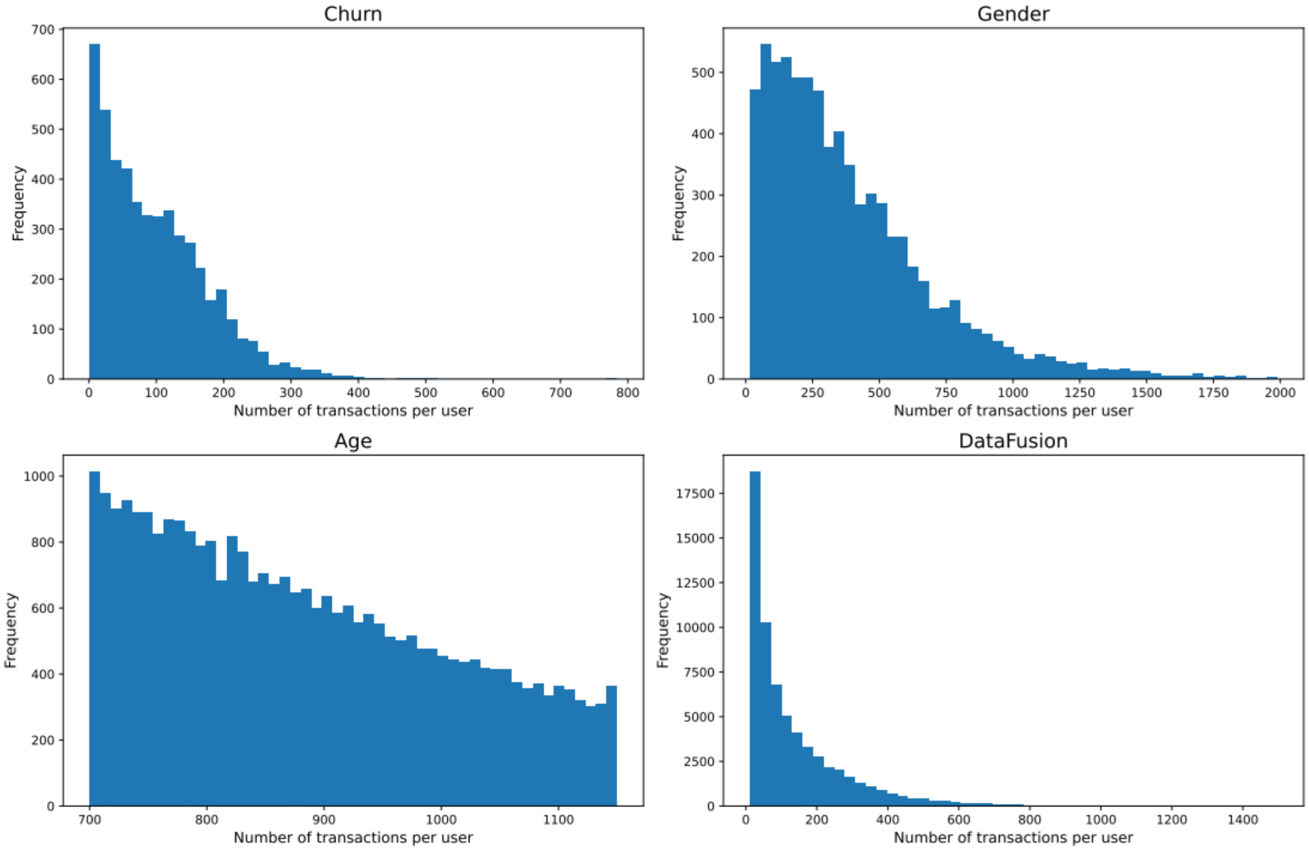
Skoltech

EDA: Amount



Skoltech

EDA: Sequence length



Skoltech