ABSTRACT

of the PhD thesis by Tokhtakhunov Il'murat Turdymagametovich
on «Deep Learning Models and Methods for Finding "Similar Audience" in Targeted Advertising»,
submitted for the degree of Doctor of Philosophy (PhD) in the EP 8D06105 – Data Science

General characteristics of the research

This thesis investigates and practically implements nonlinear dimensionality reduction and representation learning methods for high-dimensional tabular user data, with specific application to look-alike audience detection in targeted advertising systems. The study examines classical linear methods (PCA), nonlinear manifold learning approaches (t-SNE), and deep learning–based representation learning using stacked autoencoders. A stacked autoencoder framework is proposed that compresses heterogeneous user feature vectors into compact 288-dimensional latent embeddings, enabling efficient user similarity analysis without task-specific model retraining. The framework is evaluated on a large-scale anonymized telecommunications dataset comprising approximately 900,000 subscribers and 948 features organized into six distinct behavioral entity domains.

Relevance of the research

The volume of industrial tabular data is rapidly increasing across telecommunications, finance, and digital marketing. Targeted advertising systems rely on large-scale behavioral datasets with high dimensionality and heterogeneous feature spaces. Traditional machine learning pipelines struggle to scale with dynamic user behavior and continuously evolving data distributions.

Existing look-alike audience detection systems are typically based on campaign-specific classifiers that require repeated training and manual feature engineering. High-dimensional tabular data lacks a unified representation space suitable for reliable similarity modeling. Static similarity metrics applied to raw features often fail due to heterogeneous data geometry. There is a need for deep learning models capable of learning transferable representations and adaptive similarity mechanisms.

Previous studies have demonstrated interest in autoencoder-based representation learning for tabular data, but their systematic application to heterogeneous telecommunications user data for look-alike audience modeling, combined with multi-entity embedding strategies and comprehensive evaluation against both classical methods and real business metrics, has not been sufficiently studied. This gap predetermines the topic of the present dissertation and its scientific objectives.

Goal of the research

To develop models and methods of deep learning for learning universal representations and similarity mechanisms for identifying "lookalike audiences" in large-scale heterogeneous tabular data used in targeted advertising systems.

Research objectives

Objective 1. Analysis of data representation techniques for semantic capture: to analyze the applicability of existing data representation techniques for capturing semantic structure in the transformed feature space of high-dimensional heterogeneous tabular data.

Objective 2. Deep learning model for non-linear dimensionality reduction: to develop a novel deep learning model for non-linear dimensionality reduction of heterogeneous tabular data, ensuring theoretical justification of the custom decision.

Objective 3. Learned similarity mechanisms in latent embedding space: to investigate learned similarity mechanisms in latent embedding space as an alternative to classification-based look-alike audience detection, and to evaluate their effectiveness on real advertising campaigns.

Objective 4. Scalable production system and cross-domain applicability: to implement the proposed framework as a scalable production system and assess its practical applicability across multiple business domains.

Object and subject of research

Object of research: large-scale heterogeneous tabular datasets describing user behavior, device characteristics, and contextual features used in targeted advertising platforms.

Subject of research: deep learning models and methods for representation learning and similarity modeling, including multi-entity autoencoder embeddings and Siamese neural networks for look-alike audience detection.

Research methods

The research employs the following methods and tools:

- Dimensionality reduction methods: PCA (linear baseline), t-SNE (nonlinear manifold learning), stacked autoencoder (deep representation learning);
- Representation framework: multi-entity embedding strategy - six behavioral domains concatenated into a unified 288-dimensional latent space;
- Similarity learning: contrastive loss–based Siamese network trained on positive/negative user pairs from real advertising campaigns;
- Downstream evaluation: kNN classification averaged across 17 independent campaign datasets;
- Evaluation metrics: Precision, Recall, F1-score, ROC AUC, Lift Top-1, Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Score;
- Implementation stack: Python, PyTorch, PyTorch Lightning, scikit-learn, Apache Spark, MLflow, DVC.

Information base

The study is based on a large-scale anonymized dataset obtained through a collaboration agreement with a telecommunications operator. The dataset comprises records of approximately 900,000 subscribers, with 948 features derived from 2,814 raw attributes after preprocessing. Features are organized into six entity domains: User Entity (~280 features: activity, ARPU, demographics, geography), Web Entity (~190 features: internet activity and interests), Finance Entity (~80 features: anonymized transaction patterns), Device Entity (~120 features: hardware specifications and price tiers), Cell Base Station Entity (~180 features: network quality and geoactivity), and Tariff Plan Entity (~98 features: service plan parameters and pricing). Validation was performed on 17 independent advertising campaign datasets with external target variables not present during training, ensuring unbiased performance estimation and eliminating target leakage.

Correspondence to scientific development directions and state programs

The research results are consistent with the following strategic documents of the Republic of Kazakhstan:

1. Concept for the Development of Artificial Intelligence for 2024–2029 (Resolution of the Government of the Republic of Kazakhstan No. 592 of 2024) - the dissertation addresses the scientific research priority direction defined by the Concept and contributes to the development of national AI competencies in the field of machine learning and representation learning for large-scale data;
2. Address of the President of the Republic of Kazakhstan «Kazakhstan in the Era of Artificial Intelligence: Current Challenges and Their Solutions through Digital Transformation» (September 8, 2025) and the National Action Plan for its implementation - the dissertation supports the strategic objective of AI-driven digital transformation in telecommunications and digital advertising sectors;
3. Law of the Republic of Kazakhstan «On Artificial Intelligence» (No. 230-VIII, November 17, 2025) - the dissertation promotes the development and deployment of AI systems in priority sectors of the economy, in compliance with the legal framework for AI governance established by the Law;
4. National Development Plan of the Republic of Kazakhstan until 2025 and the Concept of Digital Transformation, Development of the Information and Communication Technologies sector and Cybersecurity for 2023–2029 - the dissertation contributes to data-driven digitalization of the telecommunications sector, aligned with the national priorities of digital economy development.

Scientific novelty

1. A comparative analysis of PCA, t-SNE, and autoencoder-based representation learning on large-scale heterogeneous tabular data is conducted. The autoencoder demonstrates superior semantic separation across all evaluated metrics (Silhouette Score 0.48 vs 0.21 for PCA; Davies–Bouldin Index 0.79 vs 1.84; kNN F1 0.71 vs 0.56).
2. A novel custom stacked autoencoder (948→1000→1000→288→1000→1000→948) with Batch Normalization and LeakyReLU (α=0.2) is proposed - the first architecture of this type designed for high-dimensional heterogeneous telecom tabular data. A Lemma on backpropagation is formally proved for the proposed 7-layer architecture with mixed BatchNorm–LeakyReLU activations.
3. A similarity-driven paradigm is developed integrating multi-entity embeddings (6 behavioral domains → 288-dimensional latent space) and Siamese neural networks with cosine similarity as adaptive learned similarity functions - replacing campaign-specific classifiers with a single transferable model and achieving Lift Top 1 = 12.9, ROC AUC = 0.79, Conversion Rate = 0.36 across 17 real advertising campaigns.
4. A modular production architecture is proposed and deployed (HDFS + DVC + MLflow + MinIO) processing 900,000 subscribers with monthly batch refresh in ~45 minutes and on-demand inference in ~3 minutes, confirming practical feasibility for industrial deployment. Cross-domain applicability is demonstrated for telecom, e-commerce, healthcare, and finance.

Scientific results obtained during the dissertation research

- A comprehensive comparative analysis of PCA, t-SNE, and autoencoder-based dimensionality reduction was conducted on a large-scale anonymized telecommunications dataset; the autoencoder outperforms both methods on all five evaluated metrics.
- A stacked autoencoder with architecture 948→1000→1000→288→1000→1000→948 (~5.3M parameters) was designed, trained, and validated, achieving stable convergence after 400 epochs (early stopping patience=20) with validation MSE = 0.61; a Lemma on gradient propagation through the 7-layer BatchNorm–LeakyReLU architecture was formally proved.
- A multi-entity embedding strategy integrating six behavioral domains was developed, achieving Lift Top 1 = 11.7, ROC AUC = 0.76, and Conversion Rate = 0.31 - a 77.3% improvement in Lift over the best traditional classifier (LightGBM, Lift = 6.6) and 60% over single-entity cosine similarity (Lift = 7.3).
- Siamese networks operating on precomputed autoencoder embeddings achieved Lift Top 1 = 12.9, F1 = 0.75, ROC AUC = 0.79, and Conversion Rate = 0.36, outperforming all baseline methods by an average of 41.6% across 17 independent real advertising campaigns; per-campaign conversions increased from ~17,100 (LightGBM) to ~32,400 (Siamese).
- A modular production system was designed and experimentally validated, supporting ~900,000 subscribers, monthly batch refresh in ~45 min on an 8-node cluster, and on-demand inference in ~3 min per query.

Main provisions put forward for defense

1. Multi-entity deep learning framework for effective representation learning of heterogeneous tabular data, outperforming linear and non-parametric methods on all clustering and classification quality metrics: Silhouette Score 0.48 vs 0.21 (PCA); Davies–Bouldin Index 0.79 vs 1.84; kNN F1 0.71 vs 0.56; Calinski–Harabasz Score 1020 vs 410.
2. Autoencoder-based latent embeddings (948→1000→1000→288, ~5.3M parameters) form a structured representation space that captures behavioral and contextual user characteristics, with stable convergence confirmed at 400 epochs (MSE = 0.61) and theoretical justification provided by a formally proved Lemma on gradient propagation through the BatchNorm–LeakyReLU 7-layer architecture.
3. Siamese neural network with cosine similarity operating on multi-entity autoencoder embeddings as an adaptive learned similarity function, outperforming campaign-specific

classifiers across all metrics: Lift Top 1 = 12.9, F1 = 0.75, ROC AUC = 0.79, CR = 0.36 (avg. +41.6% over SVM, Random Forest, LightGBM across 17 real campaigns).

Theoretical significance of the research

This dissertation advances the theoretical understanding of deep representation learning applied to high-dimensional heterogeneous tabular data. The formal proof of a Lemma on backpropagation through a deep stacked autoencoder with Batch Normalization and LeakyReLU activations constitutes an original theoretical contribution, providing rigorous justification for the proposed custom architecture. The systematic comparative framework for evaluating PCA, t-SNE, and autoencoder-based dimensionality reduction methods, the formalization of the multi-entity embedding strategy for unified representation of heterogeneous data, and the theoretical grounding of the similarity paradigm as an alternative to campaign-specific classifiers collectively contribute novel methodological components to the field of representation learning for tabular data.

Practical significance of the research

The research proposes and deploys a fully automated generalized look-alike service that eliminates the need for campaign-specific model retraining. The production system, operating at a telecommunications operator, serves B2B clients across 17 validated real advertising campaigns. Conversion Rate improved from 0.19 (LightGBM) to 0.36 (Siamese network) - an 89% improvement - with per-campaign conversions increasing from ~17,100 to ~32,400 across a 90,000-subscriber pool, at zero additional cost. Analyst effort is fully eliminated: no manual feature engineering or per-campaign model training is required. The methodology is domain-agnostic and applicable to e-commerce (lookalike buyer targeting), healthcare (similar patient profiling), and finance (customer segmentation from transaction embeddings). Copyright certificate No. 69310 is registered in the Republic of Kazakhstan.

Dissemination, publications and doctoral candidate's contribution to each publication

The main findings were presented and discussed at the Department of Mathematical and Computer Modelling at the International Information Technology University (2022–2026) and at the School of Digital Technologies, Narxoz University (2025–2026). A total of 5 publications were produced on the dissertation topic:

1. Tokhtakhunov I., Nurtas M., Neftissov A., Pirnaev S., Kazambayev I., Kirichenko L. Exploring Autoencoder-Based Representations for Tabular Data Classification. Engineered Science. 2025, vol. 37, 1703. DOI: 10.30919/es1703. [Scopus Q1, CiteScore: 10.2, Percentile: 88%] - The doctoral candidate independently formulated the research concept, designed and implemented the stacked autoencoder architecture, conducted all experiments on the telecommunications dataset, performed statistical analysis, and prepared the manuscript. Co-authors contributed to discussion of results and editing.
2. Tokhtakhunov I., Altaibek A., Nurtas M. Optimizing Similar Audience Search in Targeted Advertising: Effectiveness of Siamese Networks for Autoencoder-Based User Embeddings. Engineering, Technology & Applied Science Research. 2025, vol. 15, no. 3, pp. 23367–23375. DOI: 10.48084/etasr.10527. [Scopus Q2, CiteScore: 2.8, Percentile: 56%] - The doctoral candidate independently designed the Siamese network architecture, implemented the training procedure and contrastive loss framework, conducted evaluation across 17 real advertising campaigns, and wrote the manuscript. Co-authors provided consultation on advertising metrics and reviewed the manuscript.
3. Tokhtakhunov I., Nurtas M., Altaibek A., Kozhamzharova D., Aitimov M. The Efficacy of Autoencoders in the Utilization of Tabular Data for Classification Tasks. Procedia Computer Science. 2024, vol. 238, pp. 492–502. DOI: 10.1016/j.procs.2024.06.052. [Scopus Q2, CiteScore: 3.6, Percentile: 62%] - The doctoral candidate led the experimental design, implemented all baseline and autoencoder models, conducted kNN evaluation, and prepared the manuscript. Co-authors contributed to data preprocessing discussion and manuscript review.

4. Tokhtakhunov I., Nurtas M. Nonlinear Dimensionality Reduction for Lookalike Audience Detection: From Manifold Learning to Autoencoder-Based Representations. Journal of Problems in Computer Science and Information Technologies. 2026, 4(1). DOI: 10.26577/jpcsit4120268. - The doctoral candidate independently conducted the comparative analysis of PCA, t-SNE, and autoencoder methods, formalized the theoretical framework, and authored the manuscript in full. The scientific supervisor reviewed and approved the manuscript.
5. Copyright certificate No. 69310 of the Republic of Kazakhstan. Deep Learning Software Model for Look-alike Audience Discovery in Targeted Advertising Systems / Tokhtakhunov I., Nurtas M. Application 29.03.2026; Publ. 31.03.2026. - The doctoral candidate designed and implemented the complete software system, including the autoencoder training pipeline, multi-entity embedding module, Siamese network inference module, and the modular production architecture (HDFS + DVC + MLflow + MinIO). The scientific supervisor co-authored the certificate application.

Chapter overview

Chapter 1 provides a comprehensive literature review covering the curse of dimensionality, classical and nonlinear dimensionality reduction methods (PCA, t-SNE, LLE, Isomap, UMAP), autoencoder architectures for representation learning, look-alike audience modeling approaches, and Siamese network methods. The chapter identifies the research gap and formulates the theoretical basis for the proposed approach.

Chapter 2 describes the dataset (approximately 900,000 subscribers, 948 features derived from 2,814 raw attributes across six entity domains) and the complete preprocessing pipeline: one-hot encoding, missing value imputation (MICE), outlier detection using Isolation Forest, multicollinearity reduction (threshold 0.77), and min-max feature normalization.

Chapter 3 presents the mathematical formulation of all dimensionality reduction methods, including the proposed stacked autoencoder architecture (948→1000→1000→288→1000→1000→948), the formally proved Lemma on gradient propagation, the multi-entity embedding strategy, and cosine similarity computation in the latent space.

Chapter 4 describes the Siamese network architecture ([256, 512, 224, 160, 128] weight-sharing branches) designed for adaptive similarity learning on autoencoder embeddings, including the pair construction procedure using 17 campaign seed audiences and computational trade-off analysis.

Chapter 5 presents the modular production system architecture: HDFS distributed storage, DVC version control, MLflow experiment tracking, MinIO artifact storage, and dual inference modes (distributed Spark-based monthly batch refresh in ~45 min; local Pandas inference in ~3 min).

Chapter 6 presents all experimental results: visual and quantitative comparison of dimensionality reduction methods, clustering quality evaluation (Table 3), look-alike audience detection performance across 17 campaigns (Tables 4–5), Siamese network confusion matrix analysis (Table 6). The Discussion section contextualizes the findings, analyzes limitations, and outlines future research directions. The Conclusion confirms all four provisions submitted for defense.

Author's personal contribution

All key results described in the dissertation were completed by the author independently. The author was responsible for the development of the research concept; dataset acquisition and preprocessing pipeline design; design and implementation of the stacked autoencoder architecture and formal proof of the gradient propagation Lemma; development of the multi-entity embedding strategy; design and training of the Siamese network; software implementation of the complete production system (HDFS + DVC + MLflow + MinIO); experimental evaluation across 17 real advertising campaigns; statistical analysis; interpretation of results; and preparation of all manuscripts. In all publications related to the dissertation, the author played the leading role.

Structure and volume of the thesis

The dissertation includes an introduction, six main chapters, a discussion, a conclusion, a list of references, and an appendix. The total length of the main text is 98 pages, excluding appendices. The dissertation contains 17 figures, 5 tables, and 97 bibliographic references.