АННОТАЦИЯ

диссертационной работы Тохтахунова Ильмурата Турдымагаметовича «Модели и методы глубокого обучения для поиска "похожей аудитории" при таргетированной рекламе», представленной на соискание степени доктора философии (PhD) по образовательной программе: 8D06105 – «Наука о данных»

Общая характеристика исследования

Диссертация посвящена исследованию и практической реализации методов нелинейного снижения размерности и обучения представлений для высокоразмерных табличных данных о пользователях, с конкретным применением к задаче поиска похожей аудитории в системах таргетированной рекламы. В работе рассматриваются классические линейные методы (PCA), нелинейные подходы многообразного обучения (t-SNE) и обучение представлений на основе глубоких нейронных сетей с использованием стековых автоэнкодеров. Предложен фреймворк на основе стекового автоэнкодера, сжимающего разнородные векторы признаков пользователей в компактные 288-мерные латентные эмбеддинги, что обеспечивает эффективный анализ сходства пользователей без необходимости переобучения модели под конкретные задачи. Фреймворк оценён на масштабном анонимизированном датасете телекоммуникационного оператора, включающем около 900 000 абонентов и 948 признаков, объединённых в шесть поведенческих доменов.

Актуальность исследования

Объём промышленных табличных данных стремительно растёт в телекоммуникациях, финансах и digital-маркетинге. Системы таргетированной рекламы опираются на масштабные поведенческие датасеты высокой размерности с разнородными пространствами признаков. Традиционные конвейеры машинного обучения испытывают трудности с масштабированием в условиях динамичного поведения пользователей и непрерывно меняющихся распределений данных.

Существующие системы поиска похожей аудитории, как правило, основаны на специфических для каждой кампании классификаторах, требующих повторного обучения и ручного конструирования признаков. Высокоразмерные табличные данные не имеют единого пространства представлений, пригодного для надёжного моделирования сходства. Статические метрики сходства, применяемые к исходным признакам, часто дают сбои из-за разнородной геометрии данных. Требуются модели глубокого обучения, способные обучать переносимые представления и адаптивные механизмы сходства.

Предыдущие исследования демонстрируют интерес к обучению представлений на основе автоэнкодеров для табличных данных, однако их систематическое применение к разнородным данным телекоммуникационных пользователей для моделирования похожей аудитории в сочетании со стратегиями многосущностных эмбеддингов и комплексной оценкой как относительно классических методов, так и реальных бизнес-метрик остаётся недостаточно изученным. Данный пробел предопределяет тему настоящей диссертации и её научные задачи.

Цель исследования

Разработать модели и методы глубокого обучения для обучения универсальных представлений и механизмов сходства, предназначенных для идентификации “похожей аудитории” в масштабных разнородных табличных данных, используемых в системах таргетированной рекламы.

Задачи исследования

Задача 1. Анализ методов представления данных для захвата семантики: проанализировать применимость существующих методов представления данных для захвата семантической структуры в преобразованном пространстве признаков высокоразмерных разнородных табличных данных.

Задача 2. Модель глубокого обучения для нелинейного снижения размерности: разработать новую модель глубокого обучения для нелинейного снижения размерности разнородных табличных данных с теоретическим обоснованием предлагаемого решения.

Задача 3. Обученные механизмы сходства в пространстве латентных эмбеддингов: исследовать обученные механизмы сходства в пространстве латентных эмбеддингов как альтернативу поиску похожей аудитории на основе классификации и оценить их эффективность на реальных рекламных кампаниях.

Задача 4. Масштабируемая производственная система и межотраслевая применимость: реализовать предложенный фреймворк в виде масштабируемой производственной системы и оценить его практическую применимость в различных бизнес-доменах.

Объект и предмет исследования

Объект исследования: масштабные разнородные табличные датасеты, описывающие поведение пользователей, характеристики устройств и контекстные признаки, используемые в платформах таргетированной рекламы.

Предмет исследования: модели и методы глубокого обучения для обучения представлений и моделирования сходства, включая многосущностные эмбеддинги автоэнкодера и сиамские нейронные сети для поиска похожей аудитории.

Методы исследования

В исследовании применялись следующие методы и инструменты:

- Методы снижения размерности: PCA (линейная базовая линия), t-SNE (нелинейное многообразное обучение), стековый автоэнкодер (глубокое обучение представлений);
- Фреймворк представлений: стратегия многосущностных эмбеддингов - шесть поведенческих доменов, объединённых в единое 288-мерное латентное пространство;
- Обучение сходству: сиамская сеть на основе контрастных потерь, обученная на положительных/отрицательных парах пользователей из реальных рекламных кампаний;
- Оценка качества: kNN-классификация, усреднённая по 17 независимым датасетам рекламных кампаний;
- Метрики оценки: Precision, Recall, F1-score, ROC AUC, Lift Top-1, Silhouette Score, индекс Дэвиса–Болдина, индекс Калински–Харабаша;
- Стек реализации: Python, PyTorch, PyTorch Lightning, scikit-learn, Apache Spark, MLflow, DVC.

Информационная база

Исследование основано на масштабном анонимизированном датасете, полученном в рамках соглашения о сотрудничестве с телекоммуникационным оператором. Датасет включает записи около 900 000 абонентов с 948 признаками, извлечёнными из 2 814 исходных атрибутов после предобработки. Признаки организованы в шесть предметных доменов: User Entity (~280 признаков: активность, ARPU, демография, география), Web Entity (~190 признаков: интернет-активность и интересы), Finance Entity (~80 признаков: анонимизированные транзакционные паттерны), Device Entity (~120 признаков: характеристики устройств и ценовые категории), Cell Base Station Entity (~180 признаков: качество сети и геоактивность) и Tariff Plan Entity (~98 признаков: параметры тарифного плана и ценообразование). Валидация проводилась на 17 независимых датасетах рекламных кампаний с внешними целевыми переменными, отсутствовавшими при обучении, что обеспечивает несмещённую оценку качества и исключает утечку целевой переменной.

Соответствие направлениям научного развития и государственным программам

Результаты исследования соответствуют следующим стратегическим документам Республики Казахстан:

1. Концепция развития искусственного интеллекта на 2024–2029 годы (Постановление Правительства Республики Казахстан № 592 от 2024 года) - диссертация решает приоритетное направление научных исследований, определённое концепцией, и способствует развитию национальных компетенций в области ИИ в сфере машинного обучения и обучения представлений для больших данных;
2. Послание Президента Республики Казахстан «Казахстан в эпоху искусственного интеллекта: актуальные вызовы и их решения через цифровую трансформацию» (8 сентября 2025 г.) и Национальный план действий по его реализации - диссертация поддерживает стратегическую цель цифровой трансформации на основе ИИ в секторах телекоммуникаций и цифровой рекламы;
3. Закон Республики Казахстан «Об искусственном интеллекте» (№ 230-VIII от 17 ноября 2025 г.) - диссертация содействует разработке и внедрению систем ИИ в приоритетных секторах экономики в соответствии с правовой базой регулирования ИИ, установленной Законом;
4. Национальный план развития Республики Казахстан до 2025 года и Концепция цифровой трансформации, развития сектора информационно-коммуникационных технологий и кибербезопасности на 2023–2029 годы - диссертация вносит вклад в цифровизацию телекоммуникационного сектора на основе данных в соответствии с национальными приоритетами развития цифровой экономики.

Научная новизна

1. Проведён сравнительный анализ методов обучения представлений PCA, t-SNE и на основе автоэнкодера на масштабных разнородных табличных данных. Автоэнкодер демонстрирует превосходящее семантическое разделение по всем оцениваемым метрикам (Silhouette Score 0,48 против 0,21 у PCA; индекс Дэвиса–Болдина 0,79 против 1,84; kNN F1 0,71 против 0,56).
2. Предложен новый авторский стековый автоэнкодер (948→1000→1000→288→1000→1000→948) с пакетной нормализацией и LeakyReLU (α=0,2) - первая архитектура подобного типа, разработанная для высокоразмерных разнородных телекоммуникационных табличных данных. Для предложенной 7-слойной архитектуры с комбинированными активациями BatchNorm–LeakyReLU формально доказана лемма об обратном распространении ошибки.

3. Разработана парадигма на основе сходства, интегрирующая многосущностные эмбеддинги (6 поведенческих доменов → 288-мерное латентное пространство) и сиамские нейронные сети с косинусным сходством в качестве адаптивных обученных функций сходства - замена специфических для кампании классификаторов единой переносимой моделью, обеспечивающей Lift Top 1 = 12,9, ROC AUC = 0,79, Conversion Rate = 0,36 по 17 реальным рекламным кампаниям.
4. Предложена и внедрена модульная производственная архитектура (HDFS + DVC + MLflow + MinIO), обрабатывающая 900 000 абонентов с ежемесячным пакетным обновлением за ~45 минут и инференсом по запросу за ~3 минуты, что подтверждает практическую применимость для промышленного развёртывания. Продемонстрирована межотраслевая применимость для телекоммуникаций, электронной коммерции, здравоохранения и финансов.

Научные результаты, полученные в ходе диссертационного исследования

- Проведён комплексный сравнительный анализ методов снижения размерности PCA, t-SNE и на основе автоэнкодера на масштабном анонимизированном телекоммуникационном датасете; автоэнкодер превосходит оба метода по всем пяти оцениваемым метрикам.
- Спроектирован, обучен и валидирован стековый автоэнкодер с архитектурой 948→1000→1000→288→1000→1000→948 (~5,3M параметров), достигший стабильной сходимости за 400 эпох (patience ранней остановки=20) с MSE на валидации = 0,61; формально доказана лемма о распространении градиента через 7-слойную архитектуру BatchNorm–LeakyReLU.
- Разработана стратегия многосущностных эмбеддингов, объединяющая шесть поведенческих доменов, с достижением Lift Top 1 = 11,7, ROC AUC = 0,76 и Conversion Rate = 0,31 - улучшение Lift на 77,3% по сравнению с лучшим традиционным классификатором (LightGBM, Lift = 6,6) и на 60% по сравнению с однодоменным косинусным сходством (Lift = 7,3).
- Сиамские сети, работающие на предварительно вычисленных эмбеддингах автоэнкодера, достигли Lift Top 1 = 12,9, F1 = 0,75, ROC AUC = 0,79 и Conversion Rate = 0,36, превзойдя все базовые методы в среднем на 41,6% по 17 независимым реальным рекламным кампаниям; конверсии на кампанию выросли с ~17 100 (LightGBM) до ~32 400 (Siamese).
- Спроектирована и экспериментально валидирована модульная производственная система, обеспечивающая обработку ~900 000 абонентов, ежемесячное пакетное обновление за ~45 мин на 8-узловом кластере и инференс по запросу за ~3 мин на запрос.

Основные положения, выносимые на защиту

1. Многосущностный фреймворк глубокого обучения для эффективного обучения представлений разнородных табличных данных, превосходящий линейные и непараметрические методы по всем метрикам качества кластеризации и классификации: Silhouette Score 0,48 против 0,21 (PCA); индекс Дэвиса–Болдина 0,79 против 1,84; kNN F1 0,71 против 0,56; индекс Калински–Харабаша 1020 против 410.
2. Латентные эмбеддинги на основе автоэнкодера (948→1000→1000→288, ~5,3M параметров) формируют структурированное пространство представлений, захватывающее поведенческие и контекстные характеристики пользователей, со стабильной сходимостью, подтверждённой на 400 эпохах (MSE = 0,61), и теоретическим обоснованием в виде формально доказанной леммы о распространении градиента через 7-слойную архитектуру BatchNorm–LeakyReLU.
3. Сиамская нейронная сеть с косинусным сходством, работающая на многосущностных эмбеддингах автоэнкодера в качестве адаптивной обученной функции сходства, превосходящая специфические для кампании классификаторы по всем метрикам: Lift Top 1 = 12,9, F1 = 0,75, ROC AUC = 0,79, CR = 0,36 (в среднем +41,6% над SVM, Random Forest, LightGBM по 17 реальным кампаниям).

Теоретическая значимость исследования

Диссертация углубляет теоретическое понимание глубокого обучения представлений, применяемого к высокоразмерным разнородным табличным данным. Формальное доказательство леммы об обратном распространении ошибки через глубокий стековый автоэнкодер с пакетной нормализацией и активациями LeakyReLU представляет собой оригинальный теоретический вклад, обеспечивая строгое обоснование предложенной архитектуры. Систематическая сравнительная база для оценки методов снижения размерности PCA, t-SNE и на основе автоэнкодера, формализация стратегии многосущностных эмбеддингов для унифицированного представления разнородных данных и теоретическое обоснование парадигмы сходства как альтернативы специфическим для кампании классификаторам в совокупности вносят новые методологические компоненты в область обучения представлений для табличных данных.

Практическая значимость исследования

Исследование предлагает и внедряет полностью автоматизированный обобщённый сервис поиска похожей аудитории, исключающий необходимость переобучения модели для каждой кампании. Производственная система, функционирующая у телекоммуникационного оператора, обслуживает B2B-клиентов в рамках 17 валидированных реальных рекламных кампаний. Conversion Rate вырос с 0,19 (LightGBM) до 0,36 (сиамская сеть) - улучшение на 89% - при этом конверсии на кампанию увеличились с ~17 100 до ~32 400 из пула в 90 000 абонентов, без каких-либо дополнительных затрат. Трудозатраты аналитиков полностью исключены: ручное конструирование признаков и обучение модели под каждую кампанию не требуются. Методология является доменно-независимой и применима в электронной коммерции (таргетинг похожих покупателей), здравоохранении (профилирование схожих пациентов) и финансах (сегментация клиентов на основе транзакционных эмбеддингов). Авторское свидетельство № 69310 зарегистрировано в Республике Казахстан.

Апробация, публикации и вклад соискателя в каждую публикацию

Основные результаты были представлены и обсуждены на кафедре математического и компьютерного моделирования Международного университета информационных технологий (2022–2026) и в Школе цифровых технологий Университета Нархоз (2025–2026). По теме диссертации опубликовано 5 работ:

1. Токтахунов И., Нуртас М., Нефтисов А., Пирнаев С., Казамбаев И., Кириченко Л. Исследование представлений на основе автоэнкодера для классификации табличных данных. Engineered Science. 2025, т. 37, 1703. DOI: 10.30919/es1703. [Scopus Q1, CiteScore: 10,2, Percentile: 88%] - Соискатель самостоятельно сформулировал концепцию исследования, спроектировал и реализовал архитектуру стекового автоэнкодера, провёл все эксперименты на телекоммуникационном датасете, выполнил статистический анализ и подготовил рукопись. Соавторы участвовали в обсуждении результатов и редактировании.
2. Токтахунов И., Алтайбек А., Нуртас М. Оптимизация поиска похожей аудитории в таргетированной рекламе: эффективность сиамских сетей для эмбеддингов пользователей на основе автоэнкодера. Engineering, Technology & Applied Science Research. 2025, т. 15, № 3, с. 23367–23375. DOI: 10.48084/etasr.10527. [Scopus Q2, CiteScore: 2,8, Percentile: 56%] - Соискатель самостоятельно разработал архитектуру сиамской сети, реализовал процедуру обучения и фреймворк контрастных потерь, провёл оценку на 17 реальных рекламных кампаниях и написал рукопись. Соавторы предоставляли консультации по рекламным метрикам и рецензировали рукопись.
3. Токтахунов И., Нуртас М., Алтайбек А., Кожамжарова Д., Айтимов М. Эффективность автоэнкодеров при использовании табличных данных для задач классификации. Procedia Computer Science. 2024, т. 238, с. 492–502. DOI: 10.1016/j.procs.2024.06.052. [Scopus Q2, CiteScore: 3,6, Percentile: 62%] - Соискатель руководил разработкой экспериментов, реализовал все базовые модели и модели автоэнкодера, провёл kNN-оценку и подготовил рукопись. Соавторы участвовали в обсуждении предобработки данных и рецензировании рукописи.
4. Токтахунов И., Нуртас М. Нелинейное снижение размерности для обнаружения похожей аудитории: от многообразного обучения к представлениям на основе автоэнкодера. Journal of Problems in Computer Science and Information Technologies. 2026, 4(1). DOI: 10.26577/jpcsit4120268. - Соискатель самостоятельно провёл сравнительный анализ методов PCA, t-SNE и автоэнкодера, формализовал теоретическую базу и написал рукопись в полном объёме. Научный руководитель рецензировал и утвердил рукопись.
5. Авторское свидетельство № 69310 Республики Казахстан. Программная модель глубокого обучения для поиска похожей аудитории в системах таргетированной рекламы / Токтахунов И., Нуртас М. Заявка 29.03.2026; Опубл. 31.03.2026. - Соискатель спроектировал и реализовал полную программную систему, включая конвейер обучения автоэнкодера, модуль многосущностных эмбеддингов, модуль инференса сиамской сети и модульную производственную архитектуру (HDFS + DVC + MLflow + MinIO). Научный руководитель является соавтором заявки на свидетельство.

Краткое содержание глав

Глава 1 содержит комплексный обзор литературы, охватывающий проклятие размерности, классические и нелинейные методы снижения размерности (PCA, t-SNE, LLE, Isomap, UMAP), архитектуры автоэнкодеров для обучения представлений, подходы к моделированию похожей аудитории и методы сиамских сетей. В главе определяется исследовательский пробел и формулируется теоретическая основа предложенного подхода.

Глава 2 описывает датасет (около 900 000 абонентов, 948 признаков, извлечённых из 2 814 исходных атрибутов по шести предметным доменам) и полный конвейер предобработки: однократное кодирование,

импутация пропущенных значений (MICE), обнаружение выбросов методом Isolation Forest, снижение мультиколлинеарности (порог 0,77) и min-max нормализация признаков.

Глава 3 представляет математическую формулировку всех методов снижения размерности, включая предложенную архитектуру стекового автоэнкодера (948→1000→1000→288→1000→1000→948), формально доказанную лемму о распространении градиента, стратегию многосущностных эмбеддингов и вычисление косинусного сходства в латентном пространстве.

Глава 4 описывает архитектуру сиамской сети (ветви с общими весами [256, 512, 224, 160, 128]), разработанную для адаптивного обучения сходству на эмбеддингах автоэнкодера, включая процедуру формирования пар с использованием 17 начальных аудиторий рекламных кампаний и анализ вычислительных компромиссов.

Глава 5 представляет архитектуру модульной производственной системы: распределённое хранилище HDFS, контроль версий DVC, отслеживание экспериментов MLflow, хранилище артефактов MinIO и два режима инференса (распределённое пакетное ежемесячное обновление на основе Spark за ~45 мин; локальный инференс на Pandas за ~3 мин).

Глава 6 представляет все экспериментальные результаты: визуальное и количественное сравнение методов снижения размерности, оценку качества кластеризации (таблица 3), эффективность поиска похожей аудитории по 17 кампаниям (таблицы 4–5), анализ матрицы ошибок сиамской сети (таблица 6).

Раздел «Обсуждение» контекстуализирует результаты, анализирует ограничения и намечает направления будущих исследований. Заключение подтверждает все четыре положения, выносимые на защиту.

Личный вклад автора

Все ключевые результаты, описанные в диссертации, выполнены автором самостоятельно. Автор нёс ответственность за разработку концепции исследования; получение датасета и проектирование конвейера предобработки; проектирование и реализацию архитектуры стекового автоэнкодера и формальное доказательство леммы о распространении градиента; разработку стратегии многосущностных эмбеддингов; проектирование и обучение сиамской сети; программную реализацию полной производственной системы (HDFS + DVC + MLflow + MinIO); экспериментальную оценку на 17 реальных рекламных кампаниях; статистический анализ; интерпретацию результатов и подготовку всех рукописей. Во всех публикациях, связанных с диссертацией, автор выполнял ведущую роль.

Структура и объём диссертации

Диссертация включает введение, шесть основных глав, раздел обсуждения, заключение, список использованной литературы и приложение. Общий объём основного текста составляет 98 страниц, без учёта приложений. Диссертация содержит 17 рисунков, 5 таблиц и 97 библиографических ссылок.