DOI: 10.15514/ISPRAS-2026-38(5)-17

Оценка табличного глубокого обучения в условиях временного сдвига и высокоразмерных признаковых пространств

И.В. Рубачёв, *ORCID: 0000-0001-8023-8147 <irubachev@hse.ru>*

НИУ Высшая школа экономики,
101978, Россия, г. Москва, ул. Мясницкая, д. 20.

Аннотация. Прогресс в исследованиях машинного обучения важен прежде всего тогда, когда он переносится с академических наборов задач в реальные приложения. Для табличного глубокого обучения это требует понимания того, какие протоколы оценки, используемые в академии, отражают практические сценарии внедрения. В этой статье мы анализируем существующие табличные наборы задач и выявляем две характеристики, типичные для промышленных задач, но слабо представленные в популярных наборах данных: временной сдвиг распределения и признаковые пространства, полученные с помощью сложных процессов сбора данных и генерации признаков. Первая характеристика требует временных разбиений на обучающую, валидационную и тестовую выборки, а вторая меняет число и структуру предиктивных, неинформативных и коррелированных признаков по сравнению с типичными академическими наборами данных. Чтобы изучить, насколько недавние достижения табличного глубокого обучения переносятся в такие условия, мы представляем TabReD – коллекцию из восьми табличных наборов данных промышленного уровня с временными разбиениями и высокоразмерными признаковыми пространствами. Мы тестируем широкий набор методов на новом наборе задач, включая градиентный бустинг над деревьями решений, полносвязные нейронные сети, векторные представления числовых признаков, модели на основе внимания, методы с непараметрической компонентой, продвинутые рецепты обучения и ансамбли. Результаты показывают, что временная оценка меняет ранжирование методов и их относительное качество по сравнению со случайными разбиениями; на TabReD наиболее устойчивыми оказываются простые полносвязные архитектуры с векторными представлениями числовых признаков и методы градиентного бустинга, тогда как ряд более сложных современных методов переносится менее надежно.

Ключевые слова: табличные данные; набор задач; временной сдвиг; генерация признаков; глубокое обучение.

Для цитирования: Рубачёв И.В. Оценка табличного глубокого обучения в условиях временного сдвига и высокоразмерных признаковых пространств. Труды ИСП РАН, 2026, том 38, вып. 5, с. 305-324. DOI: 10.15514/ISPRAS-2026-38(5)-17

Evaluating Tabular Deep Learning under Temporal Shift and Extensive Feature Engineering

I.V. Rubachev, *ORCID: 0000-0001-8023-8147 <irubachev@hse.ru>*

National Research University, Higher School of Economics,
20, Myasnitskaya st., Moscow, 101978, Russia.

Abstract. Progress in machine learning research is valuable when it transfers from academic benchmarks to practical applications. For tabular deep learning, this requires understanding how academic evaluation settings reflect real-world applications. In this paper, we analyze existing tabular benchmarks and identify two industrially common but underrepresented properties: temporal distribution drift and feature spaces produced by extensive data acquisition and feature engineering pipelines. The first calls for time-based train/validation/test splits, while the second changes the number and structure of predictive, uninformative, and correlated features compared with typical academic datasets. To study how recent advances in tabular deep learning transfer to these conditions, we introduce TabReD, a collection of eight industry-grade tabular datasets with time-based splits and high-dimensional engineered feature spaces. We reassess a broad range of methods, including GBDTs, MLP-like neural networks, numerical feature embeddings, attention-based and retrieval-based models, advanced training recipes, and ensembles. The results show that time-based evaluation changes method rankings and relative performance compared with random splits; on TabReD, simple MLP-like architectures with numerical embeddings and GBDTs are the most robust, while several more complex recent methods transfer less reliably.

Keywords: tabular data; benchmark; temporal shift; feature engineering; deep learning.

For citation: Rubachev I.V. Evaluating Tabular Deep Learning under Temporal Shift and Extensive Feature Engineering. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 5, pp. 305-324. DOI: 10.15514/ISPRAS-2026-38(5)-17

1. Введение

Табличные данные остаются одним из наиболее распространенных форматов данных в практических приложениях машинного обучения. Они лежат в основе кредитного скоринга, страхования, электронной коммерции, логистики, рекомендательных систем, медицинских и научных задач. В отличие от изображений и текстов, где глубокое обучение стало доминирующим подходом, в задачах на табличных данных долгое время сохранялось преимущество методов градиентного бустинга над деревьями решений (Gradient-boosted decision trees, GBDT), таких как XGBoost, LightGBM и CatBoost [1-3]. Тем не менее за последние годы область табличного глубокого обучения заметно продвинулась: были предложены специализированные архитектуры, векторные представления числовых признаков, продвинутые рецепты обучения и методы с непараметрической компонентой [4-11].

Практическая ценность этого прогресса зависит от того, насколько результаты, полученные на академических наборах задач, переносятся в реальные сценарии в приложениях. Для табличных данных этот вопрос особенно важен: наборы данных из разных областей сильно различаются по числу объектов, типам признаков, наличию временной структуры и качеству разметки. Поэтому выводы о превосходстве того или иного метода могут существенно зависеть от того, какие именно наборы данных и протоколы оценки используются. Всегда есть риск, что область частично оптимизируется под хорошо доступные академические коллекции, а не под весь спектр условий, возникающих в практических системах машинного обучения (Machine Learning, ML).

В этой работе мы обращаем внимание на две характеристики промышленных табличных задач, которые часто оказываются недостаточно представлены в стандартных академических наборах задач. Первая характеристика – временная эволюция данных. Во многих приложениях модель обучается на исторических данных и затем применяется к будущим

объектам: клиентам, заказам, маршрутам, страховым заявкам или финансовым транзакциям. При этом распределения признаков и целевой переменной могут постепенно меняться. Такая ситуация хорошо известна в финансах, рекомендательных системах, логистике и ML-инженерии [12-16]. В этих условиях случайное разбиение набора данных может давать завышенную оценку качества и менять относительное сравнение моделей. Более корректной практикой является временное разбиение на обучающую, валидационную и тестовую выборки, при котором модель обучается на прошлом, выбирается по более позднему валидационному периоду и тестируется на будущем. Однако многие популярные табличные наборы данных не содержат временных меток, необходимых для такой оценки.

Вторая характеристика – высокоразмерное признаковое пространство. Во многих промышленных ML-системах наборы данных являются результатом длительных процессов сбора данных и генерации признаков. В них используются агрегаты по истории пользователей, транзакций и событий, признаки из разных временных окон, географическая информация и другие источники [5, 17-20]. В результате такие наборы могут содержать сотни признаков, многие из которых коррелированы, частично избыточны или шумны, но при этом несут полезный предиктивный сигнал. Это отличается от многих академических коллекций, где число признаков обычно существенно меньше, а сами признаки часто уже предварительно очищены и упрощены.

До последнего времени систематическая оценка табличных методов в таких условиях была затруднена: промышленные наборы данных с временными метками и большим числом сконструированных признаков редко доступны исследовательскому сообществу, а наборы соревнований, близких к реальным приложениям, используются в академических наборах задач ограниченно. Поэтому необходима переоценка современных методов на данных, которые лучше отражают практические сценарии. В этой статье мы анализируем существующие наборы задач и представляем новый набор задач TabReD, состоящий из восьми табличных наборов данных промышленного уровня. Мы проводим систематическую эмпирическую оценку широкого набора методов: GBDT, полносвязных нейросетевых моделей (MLP), MLP с векторными представлениями числовых признаков, моделей на основе внимания, методов с непараметрической компонентой, продвинутых рецептов обучения и ансамблей.

2. Вклад

Основной вклад данной работы заключается в следующем:

- Мы провели ручной анализ 100 наборов данных из распространенных табличных наборов задач и выявили две группы проблем. Первая связана с качеством данных: утечками, синтетическим или неясным происхождением их наборов, а также использованием данных, которые не являются табличными в строгом смысле. Вторая связана с недостаточным покрытием практических сценариев: временные разбиения редко используются, а наборы данных с большим числом сконструированных признаков представлены слабо.
- Мы представляем набор задач TabReD, состоящий из восьми табличных наборов данных промышленного уровня. Все наборы имеют временные разбиения на обучающую, валидационную и тестовую выборки, а также высокоразмерные пространства признаков, сформированные в результате существенной работы по генерации признаков. Часть данных отобрана из соревнований Kaggle, близких к реальным прикладным задачам; остальные данные получены из промышленных ML-систем.
- Мы проводим масштабную эмпирическую оценку методов табличного машинного обучения на TabReD. В сравнение включены методы GBDT, линейные модели,

Random Forest, MLP, ResNet, FT-Transformer, MLP с векторными представлениями числовых признаков, DCNv2, SNN, Trompt, методы с непараметрической компонентой, улучшенные рецепты обучения и ансамбли.

- Мы показываем, что временные разбиения и высокоразмерные пространства признаков существенно меняют выводы о качестве моделей. По сравнению со случайными разбиениями меняются ранги методов и относительная разница между классами моделей. На наборах данных TabReD наиболее устойчивыми оказываются GBDT и простые полносвязные нейросетевые модели с векторными представлениями числовых признаков, тогда как часть более сложных современных методов переносится в новую постановку менее надежно.

3. Обзор литературы

Глубокое обучение на табличных данных является активно развивающейся областью. Ранние работы предлагали специализированные архитектуры для табличных данных, включая SNN [4], NODE [21], TabNet [22], DCNv2 [5] и модели на основе внимания [6, 23–25]. Последующие исследования показали, что корректно настроенные простые архитектуры, такие как многослойные полносвязные нейронные сети (MLP и ResNet), часто являются не менее сильными, чем более сложные модели [6, 26–28]. Отдельное направление связано с векторными представлениями числовых признаков [7], которые позволяют простым MLP существенно улучшать качество. Также были предложены продвинутые методологии обучения, включая предобучение и аугментации [29–31], а также методы с непараметрической компонентой, такие как TabR [8] и ModernNCA [11]. Поскольку качество таких методов обычно оценивается эмпирически, выбор наборов задач и протоколов оценки является критически важным.

Табличные наборы задач и источники данных. Традиционными источниками табличных данных являются репозитории UCI [32] и OpenML [33]. Многие популярные наборы задач табличного DL опираются именно на OpenML, поскольку он предоставляет большое число наборов данных и стандартных задач. Например, набор задач из [26] и Tabzilla [34] во многом строятся на полуавтоматическом отборе данных по метаданным наборов, таким как размер выборки, число признаков и качество базовых моделей. Такой подход удобен и масштабируем, но он не заменяет ручную проверку качества данных. Как мы показываем ниже, в автоматически собранные коллекции могут попадать наборы данных с утечками, неясным происхождением данных или вовсе не табличные данные.

Другой важный источник данных – соревнования по машинному обучению на платформе Kaggle [35]. Наборы, представленные в соревнованиях, часто ближе к практическим ML-задачам: они отражают конкретные сценарии приложений и зачастую содержат больше признаков, включая агрегаты и признаки, специально построенные участниками соревнований. Тем не менее такие наборы данных сравнительно редко используются в академических наборах задач табличного DL. TabReD частично закрывает этот пробел, используя как данные соревнований Kaggle, так и новые наборы из промышленных ML-систем.

Недавние работы также подчеркивают важность качества табличных наборов задач. TableShift [36] отмечает, что наборы данных, используемые для фактического сравнительной оценки табличных моделей, часто имеют проблемы качества. TabArena [37] развивает эту линию, предлагая курируемый и поддерживаемый набор высококачественных табличных задач. Эти усилия дополняют нашу постановку: чистые наборы задач с независимыми объектами из одного распределения важны для воспроизводимой оценки базовых возможностей моделей, тогда как TabReD фокусируется на другом аспекте – проверке устойчивости методов в условиях временной эволюции данных и высокоразмерных признаковых пространств.

Наборы задач со сдвигом распределения. Сдвиг распределения активно изучается в машинном обучении, в том числе для табличных данных. TableShift [36] и WildTab [38] предлагают наборы задач, где разбиения на обучающую и тестовую выборки отражают различия между доменами. Wild-Time [39] рассматривает временные сдвиги в разных модальностях и показывает, что качество моделей может существенно ухудшаться при переходе к будущим данным. Однако эти работы в первую очередь ориентированы на оценку устойчивости к сдвигу или методы обобщения на новые домены. Наша цель отличается: мы рассматриваем широкий набор методов именно табличного ML и DL, включая современные нейросетевые архитектуры, GBDT, векторные представления числовых признаков, методы с непараметрической компонентой и продвинутые рецепты обучения.

Временная оценка особенно важна для практических приложений. В финансах, рекомендательных системах, задачах логистики и ML-инженерии давно известно, что случайные разбиения могут приводить к завышенной оценке качества и некорректному выбору модели [12-16]. В отличие от более резких доменных сдвигов, временной сдвиг часто является постепенным, а временные метки доступны разработчику модели. Поэтому он должен учитываться не как отдельная экзотическая постановка, а как стандартная часть реалистичной оценки табличных моделей.

Генерация признаков и промышленные табличные данные. Многие реальные табличные ML-системы опираются на большие процессы генерации признаков и хранилища признаков [17-20]. Такие системы создают наборы данных, существенно отличающиеся от типичных академических коллекций: они могут содержать сотни признаков, многие из которых коррелированы, частично избыточны или шумны, но при этом несут полезный сигнал. Это важно для оценки методов, поскольку разные архитектуры по-разному реагируют на высокоразмерные коррелированные пространства. Например, модели на основе внимания могут становиться вычислительно дорогими при большом числе признаков, а методы, опирающиеся на расстояния между объектами, могут страдать от сложной геометрии признакового пространства. TabReD специально включает такие наборы, чтобы сделать этот аспект оценки явным.

4. Анализ состава существующих наборов задач

Для того чтобы понять, какие условия уже покрываются существующими наборами задач табличного машинного обучения, мы провели ручной аудит 100 уникальных наборов данных классификации и регрессии из нескольких популярных коллекций [6, 8, 26, 34, 36, 38]. Для каждого набора мы анализировали источник данных, описание задачи, размер, число признаков, наличие временной структуры, возможность построения временного разбиения, а также потенциальные проблемы качества.

При анализе мы опирались на описания наборов, исходные страницы репозиториев, метаданные OpenML, материалы соревнований и предыдущие сообщения об утечках. Наборы относились к проблемным только при наличии явного признака утечки, неясного происхождения данных или несоответствия табличной постановке.

Результаты анализа приведены в табл. 1. Мы выделяем две группы наблюдений. Первая связана с качеством данных. В распространенных академических наборах задач встречаются наборы данных с утечками, наборы синтетического или неясного происхождения, а также данные, которые не являются табличными в строгом смысле согласно категоризации [40]: например, представляют собой расплющенные изображения или однородные признаки, извлеченные из одного источника. Такие данные могут быть полезны для отдельных задач, но они слабо отражают специфику промышленных табличных приложений и могут искажать выводы о качестве методов.

Вторая группа наблюдений связана с покрытием практических сценариев. Многие реальные табличные задачи имеют временную структуру: модель обучается на исторических данных и

затем применяется к будущим объектам. Однако среди проанализированных наборов временные метки часто отсутствуют, а временные разбиения на обучающую и тестовую выборки почти никогда не используются как стандартная процедура оценки. Кроме того, большинство академических коллекций содержит относительно малое число признаков. Это контрастирует с промышленными ML-системами, где наборы часто являются результатом длительной генерации признаков и содержат сотни коррелированных, частично избыточных, но предиктивных признаков. Это различие визуально показано на рис. 1: TabReD существенно смещает оценку в сторону более крупных наборов с большим числом признаков.

Табл. 1. Ландшафт существующих наборов задач табличного машинного обучения в сравнении с TabReD.

Table 1. Overview of existing tabular machine learning benchmarks compared with TabReD.

Набор задач	# объектов (Q50)	# признаков (Q50)	Утечки данных	Синтетические или неясные	Не табличные	Временное разбиение нужно	Возможно	Используется
Why Trees	16,679	13	7 / 44	1 / 44	7 / 44	22	5	✗
Tabzilla	3,087	23	3 / 36	6 / 36	12 / 36	12	0	✗
WildTab	546,543	10	1* / 3	1 / 3	0 / 3	1	1	✗
TableShift	840,582	23	0 / 15	0 / 15	0 / 15	15	8	✗
Revisiting DL	57,909	20	1* / 10	1 / 10	0 / 10	7	1	✗
TabReD	**7,163,150****	**261**	**0 / 8**	**0 / 8**	**0 / 8**	✓	✓	✓

* исходный набор имеет каноническое OOD-разбиение, но стандартное IID-разбиение содержит временную утечку.

** медианный полный размер набора данных.

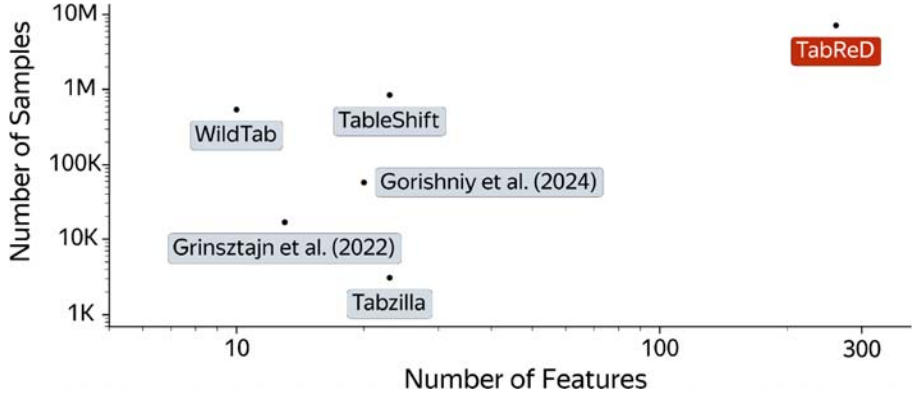

Рис. 1. Сравнение числа объектов и признаков в TabReD и существующих академических наборах задач. TabReD содержит более крупные наборы данных с большим числом признаков, что лучше соответствует ряду промышленных сценариев.

Fig. 1. Comparison of the numbers of instances and features in TabReD and existing academic benchmarks. TabReD contains larger datasets with more features, making it more representative of a range of industrial applications.

Этот анализ не означает, что существующие академические наборы задач бесполезны. Напротив, они важны для быстрой и воспроизводимой оценки моделей в относительно контролируемых условиях. Однако они не покрывают все риски практического внедрения. Для проверки переносимости методов необходим дополнительный набор задач, в котором временная эволюция данных и высокоразмерные пространства признаков являются частью основной постановки оценки.

5. Набор задач TabReD

Для устранения выявленных пробелов мы представляем TabReD – набор задач табличного машинного обучения, состоящий из восьми наборов данных промышленного уровня. Наборы TabReD отобраны так, чтобы отражать два свойства, слабо представленные в текущих академических коллекциях: временные разбиения и высокоразмерные пространства признаков, полученные в результате существенной работы по генерации признаков.

При построении TabReD мы придерживались следующих критериев. Во-первых, данные должны быть действительно табличными: строки соответствуют объектам, а столбцы – разнородным признакам, которые имеют смысл в контексте задачи. Во-вторых, признаки должны быть близки к тем, которые возникают в практических ML-системах. Для наборов из Kaggle мы адаптировали идеи и код генерации признаков из решений и обсуждений соревнований; для новых наборов использовались признаки из реальных промышленных процессов. В-третьих, мы отдельно проверяли отсутствие очевидных утечек. В-четвертых, каждый набор данных должен был содержать временные метки, достаточные для построения корректного разбиения на обучающую, валидационную и тестовую выборки по времени.

Основная информация о наборах данных приведена в табл. 2. Четыре набора взяты из соревнований Kaggle, близких к реальным прикладным задачам. Остальные четыре набора получены из промышленных ML-приложений. На рис. 2 показано компактное краткое визуальное описание задач, входящих в набор задач.

Табл. 2. Наборы данных TabReD. Числа в скобках обозначают полный размер; в экспериментах для больших наборов данных используются подвыборки.
Table 2. TabReD datasets. Numbers in parentheses indicate the full dataset size; subsets of the larger datasets are used in the experiments.

Набор данных	# объектов	# признаков	Источник	Описание задачи
Sberbank Housing	28K	392	Kaggle	прогноз цен на недвижимость
Ecom Offers	160K	119	Kaggle	предсказать, воспользуется ли пользователь скидочным предложением
Homesite Insurance	260K	299	Kaggle	прогноз принятия страхового плана
HomeCredit Default	381K (1.5M)	696	Kaggle	прогноз дефолта по кредиту
Cooking Time	319K (12.8M)	192	новые данные	оценка времени приготовления заказа
Delivery ETA	350K (17.0M)	223	новые данные	прогноз ETA курьера для доставки
Maps Routing	279K (13.6M)	986	новые данные	прогноз времени движения по маршруту
Weather	423K (16.9M)	103	новые данные	прогноз температуры

5.1 Наборы данных из соревнований Kaggle

Набор *Sberbank Housing* содержит сделки на московском рынке недвижимости; задача состоит в прогнозировании цены продажи объекта по характеристикам квартиры, района и макроэкономическим индикаторам.

Набор *Ecom Offers* моделирует лояльность пользователей в электронной коммерции: по истории транзакций нужно предсказать, воспользуется ли клиент скидочным предложением.

Набор *Homesite Insurance* ставит задачу предсказать, купит ли клиент страховой полис на дом, используя признаки клиента, полиса, продаж и географии.

Набор *HomeCredit Default* относится к банковскому скорингу: требуется предсказать дефолт клиента по кредиту на основе внутренних банковских данных и внешних источников, включая кредитные бюро.

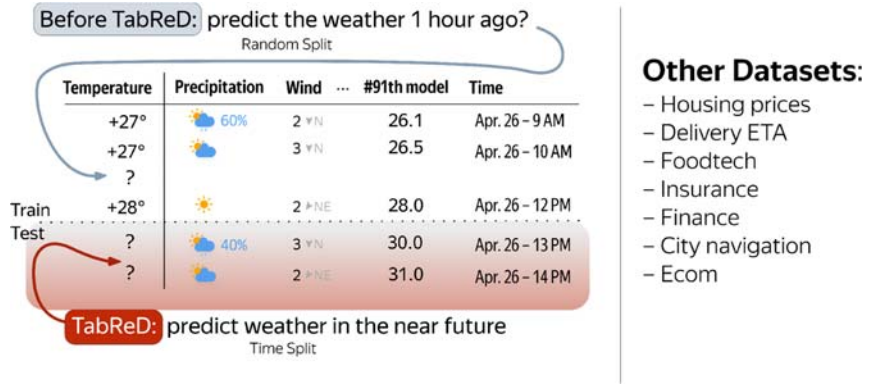

Рис. 2. Визуальное резюме задач, включенных в TabReD. Набор задач покрывает несколько прикладных доменов: недвижимость, страхование, электронную коммерцию, кредитный скоринг, доставку, маршрутизацию и прогноз погоды.
Fig. 2. Overview of the tasks included in TabReD. The benchmark covers several application areas: real estate, insurance, e-commerce, credit scoring, delivery, route planning, and weather forecasting.

5.2 Новые промышленные наборы данных

Набор *Cooking Time* оценивает время приготовления заказа рестораном в приложении доставки еды; признаки строятся по составу заказа и историческим агрегатам ресторана и бренда.

Набор *Delivery ETA* предсказывает время доставки заказа из онлайн-магазина продуктов с учетом доступности курьеров, навигационных данных и исторических временных агрегатов.

Набор *Maps Routing* прогнозирует время движения в автомобильной навигации на основе текущих дорожных условий и агрегатов по дорожному графу.

Набор *Weather* решает задачу прогноза температуры по измерениям метеостанций и физическим моделям погоды.

5.3 Свойства наборов данных TabReD

Все наборы данных TabReD имеют временные разбиения на обучающую, валидационную и тестовую выборки. Это принципиально важно: модель обучается на прошлом, подбирается по более позднему валидационному периоду и тестируется на будущих объектах. Такая постановка ближе к реальному внедрению, чем случайное IID-разбиение, особенно для задач с постепенным дрейфом распределения.

Чтобы дополнительно проверить, что временная структура наборов TabReD действительно связана со сдвигом распределения, мы сравниваем распределения обучающей и тестовой выборок при временном и случайном разбиении. В качестве простой количественной меры используем расстояние Вассерштейна: отдельно для целевой переменной и усредненно по топ-20 наиболее важных признаков. Результаты показаны на рис. 3. При временном разбиении расстояния между обучающей и тестовой выборками оказываются заметно выше, чем при случайном разбиении. Это подтверждает, что временное разбиение в TabReD не является формальностью, а действительно создает постановку со сдвигом распределения, более близкую к реальному внедрению моделей.

Второе ключевое свойство TabReD – высокоразмерные пространства признаков. Для их анализа мы строим матрицы линейных корреляций признаков и оцениваем взаимную информацию отдельных признаков с целевой переменной. Сравнение с академической коллекцией показано на рис. 4.

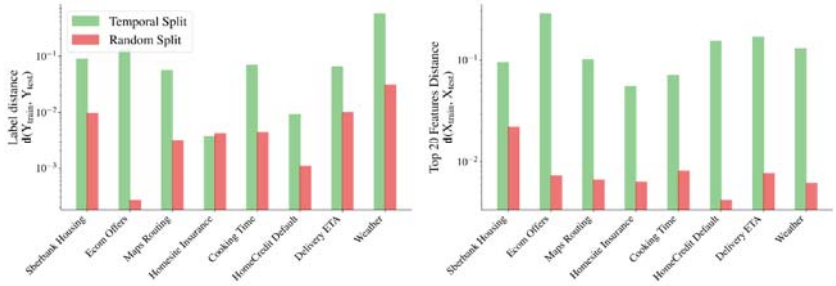

Рис. 3. Расстояние Вассерштейна между обучающей и тестовой выборками распределениями целевой переменной (слева) и топ-20 наиболее важных признаков (справа). Зеленый цвет соответствует временному разбиению, красный – случайному. При временном разбиении расстояние между обучающей и тестовой выборками выше.
Fig. 3. Wasserstein distances between the training and test distributions of the target variable (left) and the 20 most important features (right). Green denotes time-based splits; red denotes random splits. Time-based splits result in larger distances between the training and test distributions.

Наборы TabReD содержат существенно больше признаков, при этом многие из них коррелированы и имеют ненулевую связь с целевой переменной. Такая структура, естественно, возникает в результате промышленной генерации признаков: разные агрегаты и временные окна часто описывают близкие аспекты поведения объекта, но каждый из них может добавлять полезный сигнал.

TabReD следует рассматривать как дополнение к существующим наборам задач, а не как их замену. Он смещен в сторону крупных индустриально релевантных задач, где много объектов, много сконструированных признаков и присутствует временная динамика. Поэтому TabReD не покрывает все возможные домены, например медицину, социальные данные или малые научные наборы. Его роль состоит в другом: предоставить проверку того, насколько методы табличного DL сохраняют качество в условиях, которые часто возникают при практическом внедрении, но редко представлены в академической оценке.

6. Экспериментальные результаты

Далее мы используем TabReD для оценки того, какие достижения табличного глубокого обучения переносятся в условия временного сдвига и высокоразмерных пространств признаков. Экспериментальная часть устроена следующим образом. В разделе 6.3 мы сравниваем широкий набор моделей на стандартных временных разбиениях TabReD. В разделе 6.5 мы анализируем, какие именно свойства данных влияют на различия между методами: временное разбиение, размер признакового пространства и вычислительная стоимость.

6.1 Протокол оценки

Для каждого набора данных используется фиксированное временное разбиение на обучающую, валидационную и тестовую выборки. Гиперпараметры подбираются по качеству на валидационной части, после чего выбранная конфигурация оценивается на тестовой части. Для крупных наборов используются подвыборки, указанные в табл. 2; это необходимо, чтобы сделать настройку гиперпараметров большого числа методов вычислительно осуществимой.

Для оптимизации гиперпараметров используется Optuna [41] с алгоритмом Tree-structured Parzen Estimator. Для большинства методов бюджет составляет 100 итераций на пару «метод – набор данных». Исключение составляет алгоритм FT-Transformer: из-за квадратичной сложности модуля внимания по числу признаков его настройка на наборах данных с большим

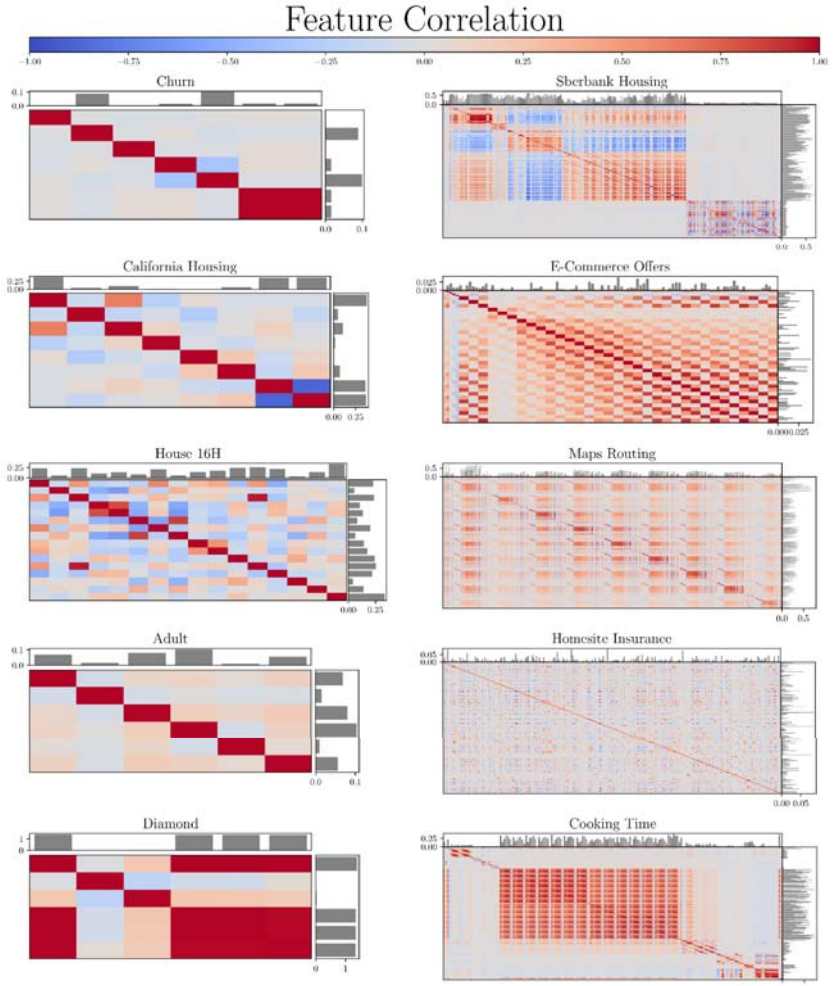

Рис. 4. Корреляции признаков и взаимная информация с целевой переменной. Слева показаны наборы данных из академической коллекции, справа – наборы TabReD. Наборы TabReD сложнее по числу признаков, структуре корреляций и паттернам важности признаков.
Fig. 4. Feature correlations and mutual information with the target variable. Datasets from the academic collection are shown on the left, and TabReD datasets on the right. TabReD datasets have greater complexity in terms of the number of features, correlation structure, and feature importance patterns.

числом признаков существенно дороже, поэтому для него используется меньший бюджет. Для Trompt используется конфигурация по умолчанию из соответствующей работы, так как полная настройка оказывается слишком дорогой.

Для MLP-подобных моделей настраивались число слоев, ширина слоев, регуляризация со случайным отключением нейронов, скорость обучения, коэффициент штрафа на норму весов и число объектов, обрабатываемых за один шаг обучения. Для GBDT настраивались глубина деревьев, скорость обучения, параметры регуляризации и параметры случайного отбора объектов и признаков. Для моделей с векторными представлениями числовых признаков

дополнительно настраивались параметры векторных представлений. Полные пространства поиска и код экспериментов доступны по запросу.

Для нейросетевых моделей используется оптимизатор AdamW [42]. В зависимости от типа задачи оптимизируется среднеквадратичная ошибка (Mean Squared Error, MSE) для регрессии или бинарная кросс-энтропия для классификации. Ранняя остановка и выбор лучшей модели выполняются по валидационному качеству. Для нейросетевых моделей числовые признаки нормализуются с помощью квантильного преобразования. Пропущенные значения заменяются средними значениями признаков после нормализации. Для категориальных признаков значения, встречающиеся только в валидационной или тестовой выборке, кодируются специальной категорией. Для регрессионных задач основной метрикой является квадратный корень из значения метрики MSE, обозначаемый как RMSE, для бинарной классификации – метрика ROC-AUC (Receiver Operating Characteristic – Area Under Curve).

Итоговые результаты усредняются по 15 случайным инициализациям. При сравнении методов мы учитываем стандартные отклонения, чтобы не интерпретировать как значимые различия, которые могут быть объяснены шумом обучения. Помимо значений метрик на отдельных наборах данных, мы используем средний ранг метода по всем задачам как компактную агрегированную характеристику качества.

6.2 Оцениваемые методы

Мы включаем несколько групп методов, чтобы покрыть основные парадигмы современного табличного ML.

Классические методы. В качестве сильных ненейросетевых базовых моделей используются три реализации градиентного бустинга над деревьями решений: XGBoost [1], LightGBM [2] и CatBoost [3]. Также оцениваются метод случайного леса (Random Forest) [43] и линейная модель. Эти методы важны как практические ориентиры: GBDT до сих пор часто является первым выбором для промышленных табличных задач.

Базовые нейросетевые модели. Мы включаем MLP как простейшую и наиболее важную базовую модель табличного DL. Также оцениваются ResNet и FT-Transformer [6]: первая представляет MLP-подобные архитектуры с остаточными соединениями, вторая – подходы на основе внимания к табличным данным.

Архитектурные варианты для табличных данных. Дополнительно оцениваются SNN [4], DCNv2 [5] и Trompt [10]. Эти методы представляют разные попытки улучшить базовые MLP-подобные архитектуры или явно моделировать взаимодействия признаков.

Векторные представления числовых признаков. Отдельно оценивается MLP-PLR – MLP с векторными представлениями числовых признаков [7]. Эта техника ранее показывала устойчивые улучшения на академических наборах задач, поэтому важно проверить, сохраняет ли она пользу в условиях TabReD.

Методы с непараметрической компонентой. Мы включаем TabR [8] и ModernNCA [11]. Эти методы используют информацию из обучающей выборки на этапе предсказания и показывали сильные результаты на стандартных академических наборах задач. В основной постановке мы оцениваем вклад непараметрической компоненты отдельно, без добавления дополнительных векторных представлений числовых признаков поверх этих моделей.

Улучшенные рецепты обучения и ансамбли. Также оцениваются методы с более длинным обучением, аугментациями и реконструкционной целью [30]. Наконец, для наиболее важных нейросетевых моделей мы рассматриваем ансамбли независимых запусков, так как ансамблирование является простой и часто эффективной практикой в табличном ML.

Таким образом, экспериментальная постановка позволяет проверить не один конкретный класс моделей, а широкий набор современных техник: от GBDT и простых MLP до архитектур на основе внимания, методов с непараметрической компонентой, векторных представлений числовых признаков, продвинутых рецептов обучения и ансамблей.

6.3 Основные результаты

Основные результаты приведены в табл. 3. Для каждого набора данных жирным выделены методы, неотличимые от лучшего с учетом стандартного отклонения по 15 случайным инициализациям. Самый правый столбец показывает усредненный по всем наборам ранг метода; меньший средний ранг означает, что метод в среднем занимает более высокое место среди сравниваемых методов.

Табл. 3. Сравнение табличных ML-моделей на наборах данных TabReD. Последний столбец содержит средний ранг алгоритма по всем наборам.
Table 3. Comparison of tabular machine learning models on the TabReD datasets. The last column reports each algorithm's average rank across all datasets.

Метод	Homesite Insurance	Ecom Offers	Home Credit Default	Sberbank Housing	Cooking Time	Delivery ETA	Map Routing	Weather	Средний ранг метода
Classical ML Baselines									
XGBoost	0.9601	0.5763	0.8670	0.2419	0.4823	0.5468	0.1616	1.4671	2.9 ± 1.5
LightGBM	0.9603	0.5758	0.8664	0.2468	0.4826	0.5468	0.1618	1.4625	3.1 ± 1.5
CatBoost	0.9606	0.5596	0.8621	0.2482	0.4823	0.5465	0.1619	1.4688	3.4 ± 1.7
RandomForest	0.9570	0.5764	0.8269	0.2640	0.4884	0.5959	0.1653	1.5838	7.8 ± 2.0
Linear	0.9290	0.5665	0.8168	0.2509	0.4882	0.5579	0.1709	1.7679	8.8 ± 2.7
Tabular DL Models									
MLP	0.9500	0.6015	0.8545	0.2508	0.4820	0.5504	0.1622	1.5470	5.0 ± 1.8
SNN	0.9492	0.5996	0.8551	0.2858	0.4838	0.5544	0.1651	1.5649	6.6 ± 2.0
DCNv2	0.9392	0.5955	0.8466	0.2770	0.4842	0.5532	0.1672	1.5782	7.6 ± 2.3
ResNet	0.9469	0.5998	0.8493	0.2743	0.4825	0.5527	0.1625	1.5021	5.8 ± 2.0
FT-Transformer	0.9622	0.5775	0.8571	0.2440	0.4820	0.5542	0.1625	1.5104	4.8 ± 1.6
MLP (PLR)	0.9621	0.5957	0.8568	0.2438	0.4812	0.5527	0.1616	1.5177	3.8 ± 1.4
Trompt	0.9588	0.5803	0.8355	0.2509	0.4809	0.5519	0.1624	1.5187	5.4 ± 2.1
Ensembles									
MLP ens.	0.9503	0.6019	0.8557	0.2447	0.4815	0.5494	0.1620	1.5186	4.1 ± 2.0
MLP-PLR ens.	0.9629	0.5981	0.8585	0.2381	0.4806	0.5518	0.1612	1.4953	2.4 ± 1.5
Training Methodologies									
MLP aug.	0.9523	0.6011	0.8449	0.2659	0.4832	0.5532	0.1631	1.5193	6.0 ± 2.0
MLP aug. rec.	0.9531	0.5960	0.7453	0.2515	0.4834	0.5541	0.1636	1.5160	6.8 ± 2.8
Retrieval Augmented Tabulard DL									
TabR	0.9487	0.5943	0.8501	0.2820	0.4828	0.5514	0.1639	1.4666	6.0 ± 2.2
ModernNCA	0.9514	0.5765	0.8531	0.2593	0.4825	0.5498	0.1625	1.5062	5.6 ± 1.6

Полученные результаты позволяют выделить несколько устойчивых закономерностей.

Методы GBDT и MLP с векторными представлениями числовых признаков являются наиболее сильными методами в среднем. На наборах данных TabReD лучшие средние результаты показывают реализации градиентного бустинга над деревьями решений и MLP-PLR. Это важное наблюдение: техника векторных представлений числовых признаков, ранее показавшая эффективность на академических наборах задач, сохраняет пользу и в более реалистичной постановке с временными разбиениями и большим числом признаков. При этом GBDT остаются очень сильной практической базовой линией, что согласуется с их широким использованием в индустриальных табличных задачах.

Ансамблирование стабильно улучшает качество. Ансамбли независимых запусков MLP и MLP-PLR дают устойчивый прирост качества. Это ожидаемый, но практически важный результат: ансамбли остаются простым и надежным способом повысить качество моделей на табличных данных, особенно когда вычислительный бюджет позволяет обучить несколько независимых моделей.

FT-Transformer конкурентоспособен, но хуже масштабируется по числу признаков. FT-Transformer показывает разумное качество и часто превосходит базовый MLP, однако его применение на TabReD осложняется вычислительной стоимостью. Механизм внимания имеет квадратичную сложность по числу признаков, что становится существенным ограничением для наборов данных с сотнями или почти тысячей признаков. В результате тщательная настройка гиперпараметров моделей на основе внимания становится заметно дороже, чем настройка MLP-подобных архитектур.

Более сложные архитектурные варианты не дают устойчивого выигрыша. SNN, DCNv2, ResNet и Trompt в среднем не превосходят базовый MLP. Это особенно показательно для Trompt: метод демонстрировал сильные результаты на некоторых академических коллекциях, однако в условиях TabReD его преимущества не переносятся надежно. Таким образом, усложнение архитектуры само по себе не гарантирует улучшения качества в данных с большим числом признаков и временной эволюцией.

Методы с непараметрической компонентой переносятся хуже, чем ожидалось. TabR и ModernNCA ранее показывали сильные результаты на стандартных академических наборах задач, однако на TabReD их качество в среднем оказывается ближе к базовому MLP и заметно уступает MLP-PLR и GBDT. Исключением является набор Weather, где TabR показывает один из лучших результатов. Это говорит о том, что поиск похожих объектов в обучающей выборке может быть полезен в отдельных задачах, но его эффективность чувствительна к структуре признакового пространства и к временному сдвигу между обучающей и тестовой выборками.

Улучшенные рецепты обучения не дают универсального выигрыша. Длительное обучение с аугментациями и дополнительной реконструкционной целью приносит пользу только на части наборов данных, например Homesite Insurance и Weather. В среднем эти методы не улучшают качество относительно базового MLP. Это отличается от результатов на более привычных академических наборах задач и указывает на ограниченную переносимость таких рецептов в новую постановку.

6.4 Сравнение с академическими наборами задач

Чтобы лучше понять, какие улучшения действительно переносятся в условия TabReD, мы сравниваем относительный прирост качества методов относительно базового MLP на двух типах наборов задач: на TabReD и на распространенной академической коллекции наборов данных, использованной в [8]. Для метрик, где большее значение лучше, относительное улучшение считается как прирост значения метрики; для RMSE используется относительное уменьшение ошибки. Результаты показаны на рис. 5.

Рис. 5. Сравнение относительных улучшений методов табличного DL на TabReD и на академическом наборе задач. Показано среднее относительное улучшение по сравнению с базовым MLP. Ансамблирование и векторные представления числовых признаков переносятся на TabReD надежно, тогда как методы с непараметрической компонентой и улучшенные рецепты обучения теряют значительную часть преимуществ.

Fig. 5. Comparison of the relative improvements achieved by tabular deep learning methods on TabReD and an academic benchmark. Mean relative improvements over the baseline MLP are shown. Ensembling and numerical feature embeddings transfer reliably to TabReD, whereas methods with a nonparametric component and enhanced training procedures lose a substantial part of their advantages.

Из сравнения видно, что не все достижения табличного DL одинаково хорошо переносятся между постановками. Ансамблирование и векторные представления числовых признаков оказываются наиболее устойчивыми: они улучшают качество как на академических наборах данных, так и на TabReD. Напротив, методы с непараметрической компонентой и улучшенные рецепты обучения показывают заметный разрыв между двумя постановками: на академическом наборе задач они дают существенный прирост относительно MLP, а на TabReD этот прирост сильно уменьшается или исчезает.

Мы предполагаем, что здесь одновременно работают два фактора. Во-первых, высокоразмерные пространства признаков TabReD содержат много коррелированных, частично избыточных и шумных признаков. Это может ухудшать качество поиска соседей и делать менее надежными аугментации, основанные на случайной замене признаков. Во-вторых, временной сдвиг ослабляет предположение о том, что объекты из обучающей выборки являются хорошими локальными аналогами для тестовых объектов. Это особенно важно для методов, которые явно используют обучающие объекты во время предсказания или склонны к запоминанию сложных примеров при длительном обучении.

Таким образом, TabReD выявляет режим, в котором часть недавних улучшений табличного DL оказывается менее универсальной, чем можно было бы заключить по академическим коллекциям. Это не отменяет их ценности, но показывает, что для оценки практической применимости методов необходимы более разнообразные наборы задач.

6.5 Влияние временных разбиений и высокоразмерных пространств признаков

Далее мы отдельно анализируем два свойства TabReD, которые отличают его от большинства академических наборов задач: временную структуру данных и большое число сконструированных признаков.

6.5.1 Временные разбиения

Для анализа влияния временного сдвига мы строим несколько временных разбиений с помощью скользящего окна по набору данных и соответствующие им случайные разбиения с теми же размерами обучающей, валидационной и тестовой выборок. Затем мы сравниваем качество методов MLP, MLP-PLR, XGBoost и TabR. Эти методы представляют разные парадигмы: параметрические нейросетевые модели, GBDT и модель с непараметрической компонентой.

Результаты показывают, что стратегия разбиения существенно влияет на выводы. Случайные разбиения часто дают более оптимистичную оценку качества, поскольку обучающие и тестовые объекты оказываются смешаны по времени. Временные разбиения, напротив, проверяют более реалистичный сценарий: модель обучается на прошлом и применяется к будущим объектам.

Важно, что меняются не только абсолютные значения метрик, но и относительное сравнение моделей. Агрегированные результаты этого сравнения приведены в табл. 4: при случайных разбиениях относительное преимущество XGBoost над MLP составляет 2.02%, тогда как при стандартных временных разбиениях TabReD оно снижается до 0.91%. В частности, преимущество XGBoost над MLP-подобными моделями уменьшается при переходе от случайных разбиений к временным. Это может означать, что часть преимущества GBDT на случайных разбиениях связана с использованием временных закономерностей, которые не переносятся на будущие данные. Такой эффект особенно важен для практики: если модель выбирается по случайному разбиению, то итоговый выбор может оказаться неоптимальным для реального внедрения.

6.5.2 Размер и структура признакового пространства

Второй анализ посвящен влиянию большого числа признаков. Для этого мы сравниваем стандартную постановку TabReD с вариантом, где оставлены только 20 наиболее важных признаков по важности XGBoost. Важности признаков вычисляются только на обучающей части соответствующего разбиения, чтобы не использовать информацию из валидационной или тестовой выборки. Такой эксперимент не является рекомендацией использовать только 20 признаков; наоборот, он позволяет понять, какие методы страдают именно от высокоразмерного, коррелированного и частично шумного пространства признаков.

Табл. 4. Влияние временного разбиения и числа признаков на относительное качество моделей. Числа показывают среднее относительное улучшение относительно MLP.
Table 4. Effect of time-based splitting and the number of features on relative model performance. Numbers indicate the mean relative improvement over MLP.

Постановка	MLP	MLP-PLR	MLP ens.	MLP-PLR ens.	XGBoost	TabR
TabReD по умолчанию	0.00%	+1.50%	+0.73%	+1.84%	+0.91%	-2.78%
Случайные разбиения	0.00%	+1.42%	+0.50%	+1.78%	+2.02%	+0.60%
Топ-20 признаков	0.00%	+1.29%	+0.29%	+1.56%	+0.97%	+1.09%

Результаты приведены в табл. 4. Во-первых, полный набор признаков действительно важен: базовый MLP со всеми признаками в среднем улучшает качество на 5.65% относительно MLP, обученного только на 20 наиболее важных признаках. Это подтверждает практическую ценность обширной генерации признаков в промышленных ML-системах.

Во-вторых, некоторые методы становятся заметно более конкурентоспособными на сокращенном признаковом пространстве. Например, TabR на полном TabReD в среднем уступает MLP, но на топ-20 признаках показывает положительное относительное улучшение. Это указывает на то, что методы с поиском соседей могут хуже работать в высокоразмерных коррелированных пространствах: расстояния между объектами становятся менее надежным источником информации, особенно при наличии временного сдвига. Похожий эффект наблюдается для методов с аугментациями: на полном признаковом пространстве они могут ухудшать качество, тогда как на топ-20 признаках вред становится меньше.

В-третьих, MLP-PLR и ансамблирование остаются полезными во всех вариантах постановки. Это делает их наиболее надежными нейросетевыми техниками среди рассмотренных: они дают прирост при случайных и временных разбиениях, а также на полном и сокращенном наборах признаков.

6.6 Практические аспекты: эффективность и устойчивость

Помимо качества, в практических табличных задачах важна вычислительная эффективность. TabReD делает этот аспект более заметным, поскольку наборы данных содержат сотни признаков и требуют настройки гиперпараметров. В таких условиях простые MLP-подобные модели имеют важное преимущество: они быстрее обучаются, проще настраиваются и легче ансамблируются. Напротив, модели на основе внимания становятся дорогими при большом числе признаков, а методы с непараметрической компонентой требуют дополнительных затрат на хранение представлений и поиск соседей.

Мы также проверяли простые методы, направленные на повышение устойчивости к распределительному сдвигу. В частности, рассматривались адаптации DeepCORAL [44], где временные интервалы трактуются как разные домены, и Deep Feature Reweighting (DFR) [45], где последняя часть обучающей выборки используется для донастройки последнего слоя модели. В наших экспериментах эти методы не дали устойчивого улучшения относительно базового MLP. Это согласуется с наблюдениями в других работах по табличным данным со сдвигом распределения: стандартные методы обобщения на новые домены не всегда переносятся на табличные задачи напрямую. Для временно эволюционирующих табличных данных, вероятно, требуются более специализированные подходы.

6.7 Итоги экспериментальной части

Эксперименты на TabReD показывают, что реалистичные свойства данных существенно меняют картину сравнения методов. На стандартных академических наборах задач многие современные техники демонстрируют большой прирост относительно MLP и GBDT. Однако в условиях временного сдвига и высокоразмерных пространств признаков этот прирост часто уменьшается или исчезает.

Наиболее надежными среди рассмотренных подходов оказываются GBDT, MLP с векторными представлениями числовых признаков и ансамбли. Более сложные архитектуры, методы с непараметрической компонентой и продвинутые рецепты обучения требуют дальнейшего анализа: их успех на академических наборах данных не всегда означает устойчивую практическую применимость. Временные разбиения при этом оказываются принципиально важны: они меняют абсолютные значения метрик, относительные отрывы между методами и итоговые ранги моделей.

7. Заключение

В этой работе мы рассмотрели проблему переносимости прогресса в табличном глубоком обучении от академических наборов задач к практическим сценариям. Мы проанализировали распространенные коллекции табличных наборов данных и выявили две группы ограничений. Первая связана с качеством данных: в используемых наборах задач встречаются утечки, наборы синтетического или неясного происхождения, а также данные, которые не являются табличными в строгом смысле. Вторая связана с покрытием практических условий: временные разбиения редко используются, а наборы с большим числом сконструированных признаков представлены слабо.

Для устранения этих пробелов мы представили TabReD – коллекцию из восьми табличных наборов данных промышленного уровня. Все наборы данных имеют временные разбиения на обучающую, валидационную и тестовую выборки и высокоразмерные пространства признаков, полученные в результате существенной работы по генерации признаков. TabReD не заменяет существующие академические наборы задач, но дополняет их: он проверяет методы в условиях, которые часто возникают при реальном внедрении ML-систем.

Эксперименты на TabReD показывают, что временные разбиения и сложные признаковые пространства существенно влияют на выводы о качестве моделей. Случайные разбиения могут давать завышенную оценку качества и менять относительные ранги методов. Многие

современные техники табличного DL, успешно работающие на академических коллекциях, переносятся в условия TabReD менее надежно. Наиболее устойчивыми в наших экспериментах оказываются GBDT, MLP с векторными представлениями числовых признаков и ансамбли, тогда как более сложные архитектуры, методы с непараметрической компонентой и улучшенные рецепты обучения показывают менее стабильный эффект.

Мы считаем, что TabReD является шагом к более репрезентативной оценке табличного машинного обучения. Дальнейшие направления включают методы работы с постепенным временным сдвигом, устойчивые методы с непараметрической компонентой, отбор и регуляризацию признаков в высокоразмерных коррелированных пространствах, а также расширение коллекции наборов данных новыми доменами. Главный вывод работы состоит в том, что прогресс в табличном DL должен оцениваться не только на удобных академических задачах, но и в постановках, отражающих реальные сложности промышленного применения.

8. Доступность данных и кода

Код подготовки данных, запуска экспериментов и построения графиков, а также инструкции по воспроизведению результатов доступны по запросу. Для новых промышленных наборов данных используются предобработанные и анонимизированные признаки; исходные внутренние журналы событий и код промышленных процессов генерации признаков не распространяются.

Список литературы / References

[1]. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. SIGKDD, 2016.

[2]. Ke G., Meng Q. et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017, vol. 30, pp. 3146-3154.

[3]. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. CatBoost: Unbiased boosting with categorical features. NeurIPS, 2018.

[4]. Klambauer G., Unterthiner T., Mayr A., Hochreiter S. Self-normalizing neural networks. NIPS, 2017.

[5]. Wang R., Shivanna R., Cheng D. Z., Jain S., Lin D., Hong L., Chi E.H. DCN V2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. arXiv preprint, 2020, DOI: 10.48550/arXiv.2008.13535v2.

[6]. Gorishniy Y., Rubachev I., Khrulkov V., Babenko A. Revisiting deep learning models for tabular data. NeurIPS, 2021.

[7]. Gorishniy Y., Rubachev I., Babenko A. On embeddings for numerical features in tabular deep learning. NeurIPS, 2022.

[8]. Gorishniy Y., Rubachev I., Kartashev N., Shlenskii D., Kotelnikov A., Babenko A. TABR: Tabular deep learning meets nearest neighbors in 2023. ICLR, 2024.

[9]. Chen J., Yan J., Chen D.Z., Wu J. ExcelFormer: A neural network surpassing GBDTs on tabular data. 2023. DOI: 10.48550/arXiv.2301.02819.

[10]. Chen K.-Y., Chiang P.-H., Chou H.-R., Chen T.-W., Chang T.-H. Trompt: Towards a better deep neural network for tabular data. ICML, 2023, vol. 202, pp. 4392-4434.

[11]. Ye H.-J., Yin H.-H., Zhan D.-C. Modern neighborhood components analysis: A deep tabular baseline two decades later. arXiv preprint, 2024. DOI: 10.48550/arXiv.2407.03257.

[12]. Stein R.M. Benchmarking default prediction models: Pitfalls and remedies in model validation. Moody's KMV, New York, 2002, vol. 20305.

[13]. Shani G. et al. Evaluating recommendation systems. Recommender systems handbook, 2011, pp. 257-297.

[14]. Herman D. et al. Home credit – credit risk model stability, 2024. Available at: https://kaggle.com/competitions/home-credit-credit-risk-model-stability, accessed 05.09.2026.

[15]. Huyen C. Designing machine learning systems. 2022. Available at: https://books.google.ru/books?id=EThwEAAAQBAJ, accessed 05.09.2026.

[16]. Baranchuk D. et al. Dedrift: Robust similarity search under content drift. Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11026-11035.

[17]. Fu Y., Soman C. Real-time data infrastructure at uber. Proceedings of the 2021 international conference on management of data, 2021, pp. 2503-2516.

[18]. Simha N. Zipline – a declarative feature engineering framework, 2020. Available at: https://youtu.be/LjcKCm0G_OY, accessed 05.09.2026.

[19]. Kakade V. ML feature serving infrastructure at lyft, 2021. Available at: https://eng.lyft.com/ml-feature-serving-infrastructure-at-lyft-d30bf2d3c32a, accessed 05.09.2026.

[20]. Anil R. et al. On the factory floor: ML engineering for industrial-scale ads recommendation models, 2022. DOI: 10.48550/arXiv.2209.05310.

[21]. Popov S., Morozov S., Babenko A. Neural oblivious decision ensembles for deep learning on tabular data. International conference on learning representations.

[22]. Arik S.O., Pfister T. TabNet: Attentive interpretable tabular learning. arXiv preprint, 2020. DOI: 10.48550/arXiv.1908.07442v5.

[23]. Song W., Shi C., Xiao Z., Duan Z., Xu Y., Zhang M., Tang J. Autoint: Automatic feature interaction learning via self-attentive neural networks. Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1161-1170.

[24]. Somepalli G. et al. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint, 2021. DOI: 10.48550/arXiv.2106.01342v1.

[25]. Kossen J., Band N., Lyle C., Gomez A. N., Rainforth T., Gal Y. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. NeurIPS, 2021.

[26]. Grinsztajn L., Oyallon E., Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? NeurIPS, the "datasets and benchmarks" track, 2022.

[27]. Shwartz-Ziv R., Armon A. Tabular data: Deep learning is not all you need. Information Fusion, 2022, vol. 81, pp. 84-90. DOI: 10.1016/j.inffus.2021.11.011.

[28]. Borisov V., Leemann T., Seßler K., Haug J., Pawelczyk M., Kasneci G. Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2024, vol. 35, no. 6, pp. 7499-7519. DOI: 10.1109/TNNLS.2022.3229161.

[29]. Bahri D., Jiang H., Tay Y., Metzler D. SCARF: Self-supervised contrastive learning using random feature corruption. ICLR, 2021.

[30]. Rubachev I., Alekberov A., Gorishniy Y., Babenko A. Revisiting pretraining objectives for tabular deep learning. arXiv preprint, 2022. DOI: 10.48550/arXiv.2207.03208.

[31]. Lee K., Sim Y.S., Cho H., Eo M., Yoon S., Yoon S., Lim W. Binning as a pretext task: Improving self-supervised learning in tabular domains. Forty-first international conference on machine learning, 2024. Available at: https://openreview.net/forum?id=ErkzxOlOLy, accessed 05.09.2026.

[32]. UCI Machine Learning Repository. University of California, Irvine. Available at: https://archive.ics.uci.edu/, accessed 05.09.2026.

[33]. OpenML. Open platform for sharing datasets and machine learning experiments. Available at: https://www.openml.org/, accessed 05.09.2026.

[34]. McElfresh D., Khandagale S., Valverde J. et al. When do neural nets outperform boosted trees on tabular data? arXiv preprint, 2023. DOI: 10.48550/arXiv.2305.02997.

[35]. Kaggle. Platform for data science competitions and datasets. Available at: https://www.kaggle.com/, accessed 05.09.2026.

[36]. Gardner J.P., Popovi Z., Schmidt L. Benchmarking distribution shift in tabular data with TableShift. Thirty-seventh conference on neural information processing systems datasets and benchmarks track, 2023.

[37]. Erickson N., Purucker L., Tschalzev A., Holzmüller D., Desai P. M., Salinas D., Hutter F. TabArena: A living benchmark for machine learning on tabular data, 2025. DOI: 10.48550/arXiv.2506.16791.

[38]. Kolesnikov S. Wild-tab: A benchmark for out-of-distribution generalization in tabular regression, 2023. DOI: 10.48550/arXiv.2312.01792.

[39]. Yao H., Choi C., Cao B., Lee Y., Koh P.W.W., Finn C. Wild-time: A benchmark of in-the-wild distribution shift over time. Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 10309-10324.

[40]. Kohli R. et al. Towards quantifying the effect of datasets for benchmarking: A look at tabular machine learning. ICLR 2024 data-centric machine learning research workshop, 2024.

[41]. Akiba T., Sano S., Yanase T., Ohta T., Koyama M. Optuna: A next-generation hyperparameter optimization framework. KDD, 2019.

[42]. Loshchilov I., Hutter F. Decoupled weight decay regularization. ICLR, 2019.

[43]. Breiman L. Random forests. Mach. Learn, 2001, vol. 45, no. 1, pp. 5-32.

[44]. Sun B., Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation. European conference on computer vision, 2016, pp. 443-450.

[45]. Kirichenko P., Izmailov P., Wilson A.G. Last layer re-training is sufficient for robustness to spurious correlations. The eleventh international conference on learning representations, 2023. Available at: https://openreview.net/forum?id=Zb6c8A-Fghk, accessed 05.09.2026.

Информация об авторах / Information about authors

Иван Викторович РУБАЧЁВ – аспирант Национального исследовательского университета «Высшая школа экономики». Сфера научных интересов: машинное обучение на табличных данных, глубокое обучение, сравнительная оценка моделей.

Ivan Viktorovich RUBACHEV – postgraduate student at HSE University. Research interests: machine learning on tabular data, deep learning, model benchmarking.