Сравнительный анализ эффективности трансформерных моделей машинного обучения и алгоритмов градиентного бустинга при решении задач классификации на клинических данных

А. Ю. Кравченко[1], А. Б. Тристанов[2], Д. А. Мнацаканян[3]

Московский институт искусственного интеллекта в здравоохранении

[1]a.u.kravchenko@yandex.ru, [2]tristanovab@zdrav.mos.ru, [3]mnatsakanyanda@zdrav.mos.ru

Аннотация. **В работе исследуется применимость табличных трансформерных моделей (TabPFN и TabDPT) при сравнении с классическими алгоритмами машинного обучения на опросных данных о состоянии здоровья и факторах риска. Эксперименты проведены на производном наборе данных CDC Diabetes Health Indicators (253 680 наблюдений, 21 признак, задача бинарной классификации) при четырёх объёмах обучающей выборки: 1 000, 10 000, 50 000 и 100 000 примеров. Основной метрикой оценки качества стал PR-AUC, учитывающий дисбаланс классов (доля положительного класса составляет 13,9%). Показано, что трансформерные модели сопоставимы с настроенными ансамблевыми методами на всех исследованных масштабах без какого-либо подбора гиперпараметров, а при больших выборках дополнительно демонстрируют высокую калибровку вероятностей. Вместе с тем по скорости инференса трансформеры значительно уступают классическим методам.**

Ключевые слова: машинное обучение; табличные данные; трансформеры; градиентный бустинг; бинарная классификация; дисбаланс классов

I. Введение

На сегодняшний день большинство прикладных задач машинного обучения связано с данными в табличном виде: строки соответствуют объектам, а столбцы признакам. Глубокое обучение за последние годы показало впечатляющие результаты в задачах распознавания изображений и работы с текстом, однако на структурированных таблицах картина иная. Здесь нейронные сети по-прежнему проигрывают ансамблевым подходам [7], а XGBoost, CatBoost и LightGBM сохраняют позиции основного рабочего инструмента как в промышленных системах, так и в соревновательной среде [5].

На этом фоне трансформерная архитектура начинает проникать и в область табличных данных. Принципиальное отличие трансформеров от классики состоит в том, что им не нужно переобучаться под каждую новую задачу: накопленные при предобучении знания переносятся на новый датасет через механизм контекстного обучения. В обработке текста такой подход уже перевернул привычный уклад, но то, насколько он применим к структурированным данным, пока остаётся вопросом без однозначного ответа.

Среди доступных решений этого класса выделяются TabPFN [1][2] и TabDPT [4]. Первый обучается целиком на синтетически сгенерированных наборах данных и строит прогноз за один прямой проход, не меняя своих весов при работе с реальными данными. TabDPT сочетает обучение в контексте с самообучением на реальных данных.

Цель настоящей работы: систематически сравнить трансформерные и классические методы при различных объёмах обучающей выборки и оценить практические границы применимости трансформеров.

II. Обзор литературы

Вопрос о конкурентоспособности нейронных сетей с деревьями решений на табличных данных активно обсуждается в литературе. Сравнительное исследование [5] показывает, что при корректной настройке гиперпараметров различия между XGBoost, CatBoost и LightGBM минимальны, а общее преимущество бустинга над нейронными архитектурами сохраняется на широком наборе датасетов. Обзор [7] охватывает десятки архитектур глубокого обучения и заключает, что ни одна из них не обеспечивает стабильного преимущества над бустингом в гетерогенных условиях.

Принципиальный сдвиг наметился с появлением подхода Prior-data Fitted Networks [1]: трансформер предобучается не на конкретном датасете, а на априорном распределении над множеством датасетов. TabPFN-2 [2], предобученный на 130 миллионах синтетических датасетов, впервые показал конкурентные с настроенными ансамблями результаты в задачах с ограниченным числом примеров. TabICL [3] масштабирует этот подход до 500 тысяч строк за счёт двухэтапной архитектуры внимания. TabDPT [4] предобучается на реальных данных и демонстрирует высокое качество без дополнительной настройки. Проблема дисбаланса классов, актуальная для медицинских и страховых данных, рассматривается в работе [6], где показано преимущество PR-AUC как метрики по сравнению с ROC-AUC при малой доле положительного класса.

III. Данные и методология

В экспериментах использовался набор CDC Diabetes Health Indicators, сформированный на основе

телефонного опроса BRFSS [8]. Набор содержит 253 680 наблюдений и 21 числовой признак. Задача: предсказать наличие сахарного диабета (бинарная классификация). Доля положительного класса составляет 13,9%, что соответствует выраженному дисбалансу. Случайный классификатор в этих условиях даёт PR-AUC около 0,139, что служит нижней границей при интерпретации результатов.

Из исходного набора методом стратифицированной выборки формировались подмножества четырёх размеров: n = 1 000, 10 000, 50 000 и 100 000 наблюдений. В каждом подмножестве данные разбивались на обучающую (70%) и тестовую (30%) части. Из обучающей части дополнительно выделялась калибровочная выборка (20%) для подбора порога классификации по критерию максимального F1 без утечки информации из тестового набора. Дубликаты (9,5% записей) сохранялись, поскольку они отражают реальное распределение опросных данных.

Сравнивались три группы моделей. Первая группа (базовые, без настройки): LogisticRegression, CatBoost, XGBoost, LightGBM, HistGradBoost, RandomForest, SVC. Вторая группа (настроенные, Optuna, 30 испытаний, 600 с на модель): CatBoost, XGBoost, LightGBM, RandomForest, SVC. Третья группа (трансформеры): TabPFN v2.5 [2] в режиме Post-Hoc Ensembling и TabDPT [4] в режиме Context Ensembling.

Метод опорных векторов (SVC) включался в эксперименты только при n не более 10 000 по причине неудовлетворительной вычислительной сложности. Алгоритм обучения SVC в стандартной реализации scikit-learn основан на решении квадратичной задачи памяти и O(n^2)–O(n^3) по времени в зависимости от ядра. Это делает SVC ресурсозатратным для больших выборок без специализированных аппроксимаций, которые выходят за рамки стандартного сравнения. Поэтому при n = 50 000 и n = 100 000 в таблице результатов для SVC указано «н/п» (не применялось).

Для TabPFN при n больше 8 000 применялось деление на подвыборки по 8 000 строк с усреднением прогнозов (4 фрагмента при n = 50 000, 7 фрагментов при n = 100 000). Для TabDPT параметры контекста подбирались под ограничения GPU (NVIDIA RTX 3070, 8,6 ГБ): от 1 024 до 2 048 строк на прогон, от 2 до 4 прогонов в зависимости от n. Для классических моделей выполнялась 5-кратная кросс-валидация (random_state = 42). Для трансформеров кросс-валидация не проводилась ввиду высоких временных затрат, что делает прямое сравнение частично асимметричным. Основная метрика: PR-AUC. Дополнительные: ROC-AUC, F1, MCC, Brier score.

IV. Результаты

В табл. 1 приведены значения PR-AUC на тестовой выборке. Звёздочкой отмечены настроенные версии алгоритмов; «н/п» означает, что модель не применялась при данном n по причинам вычислительной сложности. Нижняя граница PR-AUC для случайного классификатора на данном наборе составляет 0,139.

ТАБЛИЦА I. PR-AUC НА ТЕСТОВОЙ ВЫБОРКЕ

Модель	**1k**	**10k**	**50k**	**100k**
LogReg	0.488	0.430	0.401	0.401
CatBoost	0.414	0.411	0.416	0.413
XGBoost*	0.403	0.404	0.424	0.430
LightGBM*	0.315	0.363	0.412	0.428
SVC* (n<=10k)	0.494	0.448	н/п	н/п
TabPFN	0.462	0.428	0.432	0.432
TabDPT	0.491	0.425	0.417	0.422

При n = 1 000 все модели показали близкие результаты. Наибольший PR-AUC у SVC tuned (0,494), следом расположились TabDPT (0,491) и LogisticRegression (0,488). TabPFN достиг значения 0,462. Необходимо учитывать, что при 300 тестовых примерах точечные оценки статистически нестабильны и интерпретировать порядок моделей как устойчивый ранг без доверительных интервалов не следует. Высокий результат LogisticRegression свидетельствует о выраженной доле линейных зависимостей в данных.

С ростом n до 10 000 разрыв между группами сократился. SVC tuned показал наивысший PR-AUC (0,448), трансформеры вышли на уровень базовых бустинговых реализаций. Примечательно, что при n = 10 000 настроенный SVC всё ещё лидирует, однако его преимущество перед трансформерами уже невелико: TabPFN достиг 0,428, TabDPT – 0,425.

При n = 50 000 и n = 100 000 TabPFN показал наивысший PR-AUC в группе (0,432), немного превысив XGBoost tuned (0,424 и 0,430 соответственно). Разрыв в 0,002–0,008 не подкреплён повторными прогонами и тестами значимости, поэтому его следует трактовать как сопоставимый уровень качества, а не как доказанное превосходство. На рис. 1 показана динамика лучшего PR-AUC в каждой группе при изменении объёма выборки.

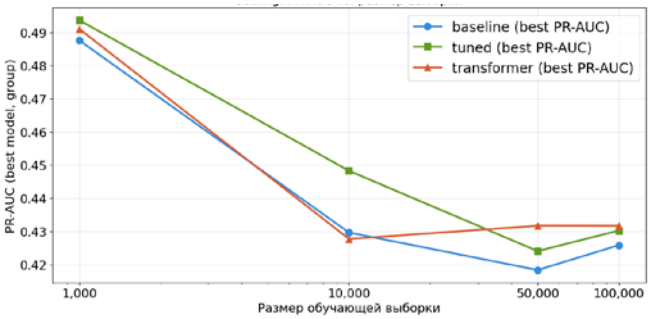

Рис. 1. PR-AUC лучших моделей по группам

Кривая трансформеров при n = 50 000 и n = 100 000 стабилизируется и незначительно превышает кривую настроенных ансамблей. Без настройки гиперпараметров трансформеры устойчиво опережают базовые классические модели при больших выборках.

В табл. 2 приведены значения Brier score и MCC для пяти ключевых моделей при n = 100 000. Меньший Brier score соответствует лучшей калибровке вероятностей.

ТАБЛИЦА II. BRIER SCORE И MCC ПРИ N = 100 000

Модель	Brier score	MCC
TabPFN	0,100	0,362
TabDPT	0,100	0,354
XGBoost tuned	0,099	0,359
CatBoost tuned	0,099	0,359
LogisticRegression	0,130	0,351

Трансформеры и настроенные бустинги показали практически одинаковый Brier score (0,099–0,100). LogisticRegression заметно уступает по данной метрике (0,130), что указывает на недостаточную откалиброванность её вероятностных прогнозов. По MCC трансформеры сопоставимы с настроенным XGBoost.

На рис. 2 показано сравнение шести метрик при n = 100 000 по всем группам.

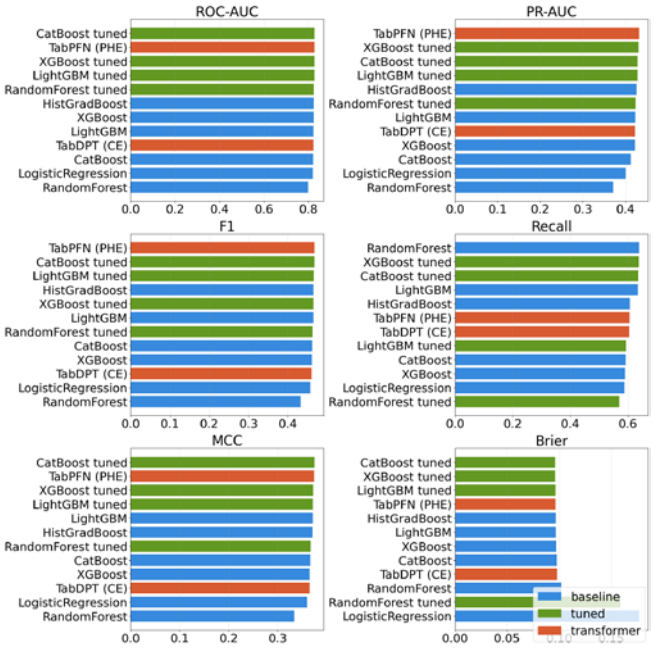

Рис. 2. Сравнение метрик при n = 100 000

Время обучения трансформеров при n = 1000 составляет около 5–25 секунд для TabPFN и 1–5 секунд для TabDPT. При n = 10 000 оно достигает 27 и 5 секунд соответственно. Время инференса TabPFN при n = 100 000 составляет около 2,3 часа, TabDPT — около 6 часов. Настроенный XGBoost при том же объёме формирует прогноз менее чем за одну секунду. В производственных системах с требованиями к задержке это принципиальное различие.

V. ОБСУЖДЕНИЕ

Полученные результаты следует рассматривать с учётом нескольких ограничений. Эксперименты выполнены на одном наборе данных, который имеет достаточно конкретные свойства: числовые признаки, опросный характер данных, бинарная постановка задачи и выраженный дисбаланс классов. Поэтому переносить сделанные выводы на регрессионные задачи, многоклассовую классификацию или другие предметные области без дополнительной проверки преждевременно. Стоит также учитывать, что сравнение трансформеров с классическими моделями по точечным значениям PR-AUC нельзя назвать полностью равноправным. Классические алгоритмы проходили через процедуры настройки гиперпараметров и кросс-валидацию, тогда как для трансформеров кросс-валидация не проводилась из-за чрезмерно высоких вычислительных затрат.

При этом в результатах прослеживается закономерность, которая имеет практическое значение. Трансформеры выходят на конкурентный уровень качества без какого-либо подбора параметров. Это особенно актуально тогда, когда ресурс на оптимизацию моделей ограничен или нужно быстро получить рабочую базовую модель. На малых выборках, при n не более 10 000, TabDPT и TabPFN представляют собой вполне разумную отправную точку. По большинству метрик они держатся на уровне базовых бустинговых реализаций и при этом не требуют от исследователя никакого ручного перебора параметров.

Отдельно стоит остановиться на качестве вероятностных оценок. Есть задачи, где на выходе модели нужна не просто метка, а числовая оценка риска: медицинский скрининг, ранняя диагностика, системы поддержки решений. В таких случаях калибровка становится самостоятельным критерием качества. При n = 100 000 оба трансформера показали Brier score на уровне настроенных бустинговых моделей. Базовая логистическая регрессия при этом заметно отстала. Из этого не следует, что трансформеры лучше во всех ситуациях, однако там, где важна точность вероятностного прогноза, они оказываются не просто альтернативой, а действительно подходящим выбором.

VI. ЗАКЛЮЧЕНИЕ

Проведённое сравнение показало, что табличные трансформеры TabPFN и TabDPT способны достигать качества, сопоставимого с настроенными ансамблевыми методами, по основным метрикам: PR-AUC, MCC и Brier score. Это наблюдалось на всех исследованных объёмах обучающей выборки. Важным практическим преимуществом трансформеров является отсутствие необходимости в подборе гиперпараметров: в отличие от бустинговых моделей, они позволяют получить конкурентный результат без дополнительного тюнинга. Для прикладных задач это может быть существенным фактором, особенно если вычислительные и временные ресурсы на настройку ограничены.

На больших выборках, начиная с n = 50 000, TabPFN показал несколько более высокий PR-AUC, чем настроенный XGBoost. Тем не менее, интерпретировать этот результат нужно осторожно. Наблюдаемый разрыв невелик, повторные прогоны не проводились, статистические тесты не применялись. Поэтому говорить о доказанном превосходстве преждевременно. Точнее будет сказать, что трансформеры показали сопоставимый, а местами чуть более высокий результат.

Главным практическим ограничением трансформерных моделей остаётся время инференса. При n = 100 000 предсказание занимает часы, тогда как бустинговые модели справляются за доли секунды. Это существенно ограничивает их применимость в системах с большим объём данных, но где задержка ответа критична. Более подходящими для трансформеров являются задачи с небольшими и средними выборками, когда нет времени на длительную настройку моделей, но важна хорошая калибровка вероятностей.

В дальнейшем результаты стоит проверить на других наборах данных и в разных предметных областях. Также необходимо оценить, являются ли различия между моделями статистически значимыми. Отдельно следует изучить, можно ли улучшить качество табличных трансформеров за счёт дообучения на доменных данных. Кроме того, важно подробнее рассмотреть баланс между качеством прогноза, калибровкой вероятностей и скоростью работы модели, поскольку именно он определяет, насколько такие методы пригодны для практического использования.

Список литературы

[1] Hollmann N., Müller S., Eggensperger K., Hutter F. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second // Proc. ICLR. 2023. URL: https://openreview.net/forum?id=cp5PvcI6w8_ (дата обращения: 14.04.2025).

[2] Hollmann N., Müller S., Purucker L., Krishnakumar A., Körfer M., Hoo S.B., Schirrmeister R.T., Hutter F. Accurate Predictions on Small Data with a Tabular Foundation Model // Nature. 2025. Vol. 637, № 8045. P. 319–326. DOI: 10.1038/s41586-024-08328-6.

[3] Qu J., Holzmüller D., Varoquaux G., Le Morvan M. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data // arXiv. 2025. arXiv:2502.05564. URL: https://arxiv.org/abs/2502.05564 (дата обращения: 16.04.2025).

[4] Ma J., Thomas V., Hosseinzadeh R., Labach A., Kamkari H., Cresswell J.C., Golestan K., Yu G., Volkovs M., Caterini A.L. TabDPT: Scaling Tabular Foundation Models on Real Data // arXiv. 2024. arXiv:2410.18164. DOI: 10.48550/arXiv.2410.18164. URL: https://arxiv.org/abs/2410.18164 (дата обращения: 16.04.2025).

[5] Пономарев Д.С. Сопоставление библиотек для создания моделей машинного обучения на основе методов градиентного бустинга // Совр. инновации, системы и технологии. 2025. Т. 5, № 2. С. 3001–3006. DOI: 10.47813/2782-2818-2025-5-2-3001-3006.

[6] Константинов А.Ф., Дьяконова Л.П. Сравнительный анализ методов снижения дисбаланса классов при построении моделей машинного обучения в финансовом секторе // Изв. Каб.-Балк. науч. центра РАН. 2025. Т. 27, № 1. С. 143–151. DOI: 10.35330/1991-6639-2025-27-1-143-151.

[7] Borisov V., Leemann T., Seßler K., Grabocka J., Nummert M., Keuper J., Kasneci G. Deep Neural Networks and Tabular Data: A Survey // IEEE Trans. Neural Netw. Learn. Syst. 2024. Vol. 35, № 6. P. 7499–7519. DOI: 10.1109/TNNLS.2022.3229161.

[8] CDC Diabetes Health Indicators Dataset. Kaggle. URL: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset (дата обращения: 14.04.2025).