bilimger.kz
Республикалық білім және ғылым порталы
https://bilimger.kz
ҚР Мәдениет және Ақпарат министрлігі тіркелген — Куәлік нөмірі: ***KZ45VPY00102718***
2026

БӨЛІМ: FARABI UNIVERSITY / UNIVER / СТУДЕНТ

Разработка интеллектуальной системы прогнозирования оттока клиентов на основе методов машинного обучения

ЖАРИЯЛАНДЫ	СІЛТЕМЕ
22.04.2026	https://bilimger.kz/188330/

Ким Сергей Юрьевич
Әл-Фараби атындағы Қазақ ұлттық университеті
Студент 4 курса, направление «Информационные системы»

Аннотация. В статье представлены результаты сравнительного исследования шести алгоритмов прогнозирования оттока клиентов — Logistic Regression, Random Forest, Gradient Boosting, XGBoost, TabNet и FT-Transformer — на трёх отраслевых наборах данных: банковском (10 000 наблюдений), телекоммуникационном (7 043 наблюдения) и сотовой связи (71 047 наблюдений). Описана архитектура разработанной многомодульной веб-системы на FastAPI и React с SHAP-интерпретацией прогнозов. Установлено, что FT-Transformer достигает наилучших показателей ROC-AUC среди нейросетевых архитектур (0,840 и 0,865 для Telco и Bank соответственно), конкурируя с Random Forest (0,842 и 0,863). Рассмотрен эксперимент по межотраслевому переносу моделей на датасет Cell2Cell.

Ключевые слова: отток клиентов, машинное обучение, FT-Transformer, TabNet, XGBoost, Random Forest, SHAP, ROC-AUC, FastAPI, нейронные сети.

1. ВВЕДЕНИЕ

Удержание существующих клиентов является одним из ключевых факторов конкурентоспособности компаний в современных высококонкурентных отраслях. По оценкам Bain & Company, повышение уровня удержания клиентов на 5% способно увеличить прибыль компании на 25–95% в зависимости от сферы бизнеса. Согласно данным Harvard Business Review, привлечение нового клиента обходится в 5–7 раз дороже, чем сохранение существующего [1].

Особую остроту проблема оттока приобретает в банковской сфере,

телекоммуникациях и мобильной связи, где годовой отток достигает 15–30% клиентской базы. Традиционные реактивные стратегии удержания уступают место превентивным подходам, основанным на прогностической аналитике с применением методов машинного обучения.

Целью настоящей статьи является изложение методологии и результатов разработки интеллектуальной веб-системы прогнозирования оттока клиентов, сочетающей классические алгоритмы машинного обучения и архитектуры глубокого обучения нового поколения — TabNet и FT-Transformer — с функционалом SHAP-интерпретации и межотраслевого сравнения.

2. ПОСТАНОВКА ЗАДАЧИ И ДАННЫЕ

Задача прогнозирования оттока формализуется как задача бинарной классификации с учителем: для каждого клиента xi ∈ Rd требуется предсказать целевую переменную yi ∈ {0; 1}, где 1 означает уход клиента, 0 — сохранение лояльности. Основной метрикой качества выбран ROC-AUC, устойчивый к дисбалансу классов и позволяющий сравнивать модели независимо от порогового значения.

Эмпирическую базу исследования составили три общедоступных набора данных. Банковский датасет (Bank Churn) включает 10 000 записей с долей оттока 20,4%; ключевыми предикторами выступают возраст клиента (Age) и количество банковских продуктов (NumOfProducts). Телекоммуникационный датасет (Telco Churn) содержит 7 043 наблюдения с долей оттока 26,5%; наиболее значимые признаки — тип контракта (Contract) и срок обслуживания (tenure). Датасет сотовой связи Cell2Cell насчитывает 71 047 записей и использовался исключительно в качестве holdout-выборки для тестирования межотраслевого переноса моделей. Дополнительно в систему интегрирован оригинальный датасет казахстанского рынка мобильной связи KZ Telecom Churn (70 000 записей, доля оттока 27,84%).

Конвейер предобработки данных включает стратифицированное трёхчастное разбиение в пропорции 70/15/15 (train/val/test), обработку пропущенных значений (SimpleImputer), нормализацию числовых признаков (StandardScaler) и кодирование категориальных переменных (OrdinalEncoder). Принципиально важно, что все трансформирующие преобразователи обучаются исключительно на тренировочной выборке — это исключает утечку данных (data leakage) и обеспечивает объективность финальных оценок качества на тестовой выборке.

3. ПРИМЕНЯЕМЫЕ МЕТОДЫ

3.1 Классические алгоритмы машинного обучения

В работе реализованы четыре классических алгоритма. Logistic Regression служит базовым ориентиром и обеспечивает интерпретируемость линейных зависимостей.

Random Forest (Breiman, 2001) — ансамбль из 200 деревьев решений с параметром class_weight='balanced', компенсирующим дисбаланс классов [2]. Gradient Boosting и XGBoost (Chen & Guestrin, 2016) реализуют последовательное построение ансамбля с минимизацией ошибок на псевдоостатках [3]. Для всех классических моделей применялась пятикратная стратифицированная кросс-валидация.

3.2 Нейросетевые архитектуры для табличных данных

TabNet (Arik & Pfister, AAAI 2021) реализует последовательный выбор признаков через механизм sparsemax-внимания, обеспечивая встроенную интерпретируемость через attention-маски [4]. Архитектура FT-Transformer (Gorishniy et al., NeurIPS 2021) преобразует каждый признак в токен-эмбеддинг (Feature Tokenizer), после чего применяет механизм Self-Attention по всем токенам — аналогично трансформерным архитектурам в обработке естественного языка [5]. Обе нейросети обучались с ранней остановкой по валидационному ROC-AUC; порог классификации подбирался по максимуму F1 на валидационной выборке.

3.3 Интерпретируемость: метод SHAP

Для объяснения индивидуальных прогнозов применяется метод SHAP (SHapley Additive exPlanations, Lundberg & Lee, 2017), основанный на теории кооперативных игр Шепли [6]. SHAP-значение каждого признака отражает его вклад в отклонение предсказанной вероятности оттока от среднего по датасету. Метод обладает свойством локальной точности: сумма SHAP-вкладов по всем признакам равна разности индивидуального прогноза и базового значения.

4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ

Итоговые результаты сравнительного тестирования всех шести моделей на тестовых выборках двух размеченных датасетов представлены в таблице 1.

Таблица 1 — Сводное сравнение моделей по ROC-AUC на тестовой выборке

Модель	Тип	Telco AUC	Bank AUC	Среднее
FT-Transformer	Нейросеть	0,840	0,865	0,853
Random Forest	Классич. ML	0,842	0,863	0,853
Logistic Regression	Классич. ML	0,840	0,844	0,842
XGBoost	Классич. ML	0,832	0,854	0,843
Gradient Boosting	Классич. ML	0,825	0,858	0,842
TabNet	Нейросеть	0,799	0,827	0,813

FT-Transformer занимает первое место по среднему ROC-AUC среди нейросетевых архитектур, показывая 0,840 на Telco и 0,865 на Bank — наилучший результат среди всех

шести моделей на банковском датасете. Превосходство над TabNet составляет 0,038–0,041 по ROC-AUC, что является статистически значимым. Полученные результаты согласуются с выводами оригинальной работы Gorishniy et al. (NeurIPS 2021) о конкурентоспособности трансформерных архитектур на табличных данных.

Random Forest демонстрирует исключительную стабильность: его ROC-AUC практически равен FT-Transformer (разрыв 0,002 на обоих датасетах), при этом модель обеспечивает более высокий Recall (0,783 на Telco) благодаря параметру class_weight='balanced'. Это делает Random Forest предпочтительным выбором в сценариях с высокой стоимостью пропущенного уходящего клиента.

Показательным является результат Logistic Regression на Telco (ROC-AUC = 0,840 — наравне с Random Forest и FT-Transformer). Это свидетельствует о том, что значительная часть предсказательной информации в телекоммуникационном датасете кодируется линейными зависимостями, что подтверждается высокими корреляциями Пирсона: Contract (+0,405) и tenure (−0,352). На банковском датасете линейная модель уступает ансамблевым методам значительнее (0,844 против 0,863–0,865), отражая более сложную нелинейную структуру взаимодействий признаков.

Анализ важности признаков через механизм attention-масок TabNet подтверждает различия в предикторах оттока по отраслям. Для телекоммуникационного датасета доминирующими признаками являются tenure (0,115) и MonthlyCharges (0,100). Для банковского — NumOfProducts (0,152), Balance (0,145) и Age (0,144). Совпадение рейтингов, независимо полученных методами attention-масок и MDI деревьев (Random Forest), свидетельствует о робастности этих закономерностей.

5. МЕЖОТРАСЛЕВОЙ ПЕРЕНОС МОДЕЛЕЙ

Для оценки обобщающей способности нейросетевых архитектур был проведён двухэтапный эксперимент по межотраслевому переносу на датасете Cell2Cell holdout (20 000 записей без меток).

На первом этапе (прямой перенос Telco → Cell2Cell без дообучения) TabNet продемонстрировал полную деградацию: среднее предсказанных вероятностей составило 0,064, а предсказанный churn rate при использовании val-порога $\theta = 0{,}22$ — 0%. Причина — несовместимость пространств признаков и неспособность sparsemax-механизма формировать значимые представления при массовом кодировании неизвестных категорий значением -1. FT-Transformer, напротив, сохраняет осмысленные вероятности (mean = 0,297) благодаря механизму Self-Attention, способному извлекать переносимые паттерны даже при частичной деградации токенов.

На втором этапе (обучение непосредственно на Cell2Cell) обе архитектуры генерировали бимодальные распределения предсказаний — диагностический признак

уверенного классификатора. Предсказанный churn rate при этом составил ~49–50%, что существенно выше истинного ~15%. Это расхождение указывает на необходимость калибровки вероятностей (Platt Scaling или Isotonic Regression) при кросс-доменном применении.

6. АРХИТЕКТУРА ВЕБ-СИСТЕМЫ

На основе результатов моделирования разработана полнофункциональная веб-система прогнозирования оттока клиентов. Серверная часть реализована на FastAPI (Python) с кэшированием моделей в оперативной памяти, что обеспечивает время отклика от 12 мс на повторных запросах (против 820 мс при холодном старте). Клиентская часть построена на React 18.2 с поддержкой светлой и тёмной темы.

Система структурирована в семь функциональных модулей: Dashboard с KPI-карточками и динамикой churn rate; Аналитика с корреляционным анализом признаков; Предсказания с gauge-шкалой вероятности оттока; Сравнение моделей с реальными ROC-кривыми; SHAP-интерпретация с визуализацией вкладов признаков; Сравнение отраслей по ключевым предикторам; Churn Analysis с сегментацией клиентов.

Аутентификация реализована на основе SHA-256 хэширования файла-ключа с истечением сессии через 24 часа. REST API включает 13 эндпоинтов, все проверены на корректность возвращаемых HTTP-статусов и математическую корректность SHAP-объяснений (свойство локальной точности, отклонение ±0,01).

7. ЗАКЛЮЧЕНИЕ

В ходе исследования разработана и апробирована интеллектуальная многомодульная система прогнозирования оттока клиентов, охватывающая банковскую, телекоммуникационную и сотовую отрасли, а также казахстанский рынок мобильной связи. Проведённый сравнительный анализ шести алгоритмов позволяет сформулировать следующие выводы.

FT-Transformer является лучшей или равной лучшей моделью по ROC-AUC на обоих датасетах (0,840–0,865), подтверждая применимость трансформерных архитектур к задачам прогнозирования на табличных данных. Random Forest демонстрирует сопоставимые ROC-AUC при более высоком Recall, что делает его предпочтительным при высокой стоимости пропущенных уходящих клиентов. TabNet уступает FT-Transformer по всем метрикам, однако обладает встроенной интерпретируемостью и устойчивостью к работе с низкими порогами. При межотраслевом переносе без дообучения FT-Transformer значительно превосходит TabNet по устойчивости к доменному сдвигу.

Практическая значимость работы состоит в создании готового к эксплуатации программного продукта с SHAP-интерпретацией, позволяющего аналитикам и менеджерам по работе с клиентами принимать обоснованные управленческие решения

по удержанию. Направления дальнейшего развития: калибровка вероятностей для кросс-доменного применения, интеграция LightGBM и CatBoost, автоматизированная настройка гиперпараметров через Optuna.

СПИСОК ЛИТЕРАТУРЫ

1. *Reichheld, F. F. The One Number You Need to Grow // Harvard Business Review. — 2003. — Vol. 81, No. 12. — P. 46–54.*
2. *Breiman, L. Random Forests // Machine Learning. — 2001. — Vol. 45, No. 1. — P. 5–32.*
3. *Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System // Proceedings of SIGKDD. — 2016. — P. 785–794.*
4. *Arik, S. O., Pfister, T. TabNet: Attentive Interpretable Tabular Learning // AAAI. — 2021. — Vol. 35, No. 8. — P. 6679–6687.*
5. *Gorishniy, Y. et al. Revisiting Deep Learning Models for Tabular Data // NeurIPS. — 2021. — Vol. 34. — P. 18932–18943.*
6. *Lundberg, S. M., Lee, S. A Unified Approach to Interpreting Model Predictions // NeurIPS. — 2017. — Vol. 30. — P. 4765–4774.*

ҚМ АА Куәлік нөмірі: **KZ45VPY00102718** — ҚР Мәдениет және Ақпарат министрлігі