УДК 004.852:336.77

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА КРЕДИТНОГО СКОРИНГА НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОЦЕНКИ КРЕДИТОСПОСОБНОСТИ ЗАЁМЩИКОВ

Орынбасар Ізбасар Орынбасарұлы
izok2004@gmail.com
магистрант 1-курса ОП «Программной инженерии»
Атырауский университет имени Х. Досмухамедова, г. Атырау, Республика Казахстан
Научный руководитель - PhD доктор, ассоциированный профессор,
декан факультета «Программной инженерии» Асанова Бактыгул Унгарсиновна

Введение

Кредитный скоринг – сфера, где риск совершить ошибку имеет особенно высокую стоимость. Если банк выдает кредит тому, кто оказался недобросовестным заемщиком, это ведет к прямым финансовыми потерями. Но отказавшись от сотрудничества с надежным заемщиком, он теряет потенциальные прибыли. Именно поэтому банки не прекращают долгое время работать над тем, чтобы повысить точность различения заемщиков по степени их риска [1].

Тема особенно актуальна в Казахстане, где с 2022 года действуют механизмы реструктуризации долга и банкротства частных лиц. Они выполняют важную социальную функцию, позволяющую гражданам покончить с финансовыми проблемами. Однако, реализация таких механизмов приводит к увеличению количества заявок на банкротство и списания задолженностей. Все это дополнительно нагружает систему банков [1].

Вместе с этим ухудшается качество кредитных портфелей. Увеличение доли невозвратных займов влечет за собой увеличение кредитного риска, оказывающее воздействие как на устойчивость банковской системы, так и бюджетные расходы. Таким образом, в настоящее время возникла ситуация, когда негативные последствия затрагивают все заинтересованные стороны: финансовые учреждения, государство и заемщики.

Основной проблемой остается ограниченность традиционных скоринговых моделей, построенных на основе линейных зависимостей. Такие модели не могут учитывать нелинейные взаимосвязи между признаками, которые играют важную роль в процессе кредитования. Например, общее влияние факторов такого рода как уровень дохода, занятость и кредитная активность может быть нелинейным [2, 3].

Новые методы машинного обучения автоматически находят такие взаимосвязи. При этом они представляют собой эффективное решение проблемы, что доказано на примере многочисленных исследований.

В данном исследовании проводится анализ нескольких подходов – от линейной логистической регрессии до алгоритмов градиентного бустинга, с целью определения наиболее эффективной модели для прогнозирования дефолта, а также оценки компромисса между точностью и интерпретируемостью.

Постановка проблемы

На данный момент в Казахстане банки испытывают ряд связанных друг с другом проблем из-за увеличения кредитных рисков. Во-первых, рост случаев невозврата кредитов приводит к увеличению количества проблемных активов в банках, что снижает ликвидность и делает менее доходными банки, а значит, делает их менее стабильными. При этом государству приходится компенсировать часть списываемых обязательств за счет средств, предназначенных для социально значимых сфер: образования, здравоохранения, строительства и развития инфраструктуры [1].

Затронутыми данной проблемой оказываются также простые граждане. Процесс банкротства человека, несмотря на его социальную направленность, не позволяет ему

пользоваться финансами в течение нескольких лет, негативно сказывается на его кредитной истории и не дает шансов развивать экономически выгодную деятельность.

Доступ к кредиту предоставляют скоринговые модели, основанные на логистической регрессии и ручной классификации входных переменных. Тем не менее их применение является крайне ограниченным вследствие невозможности корректного учета нелинейных зависимостей между переменными, поведенческой и транзакционной информации, количество которой постоянно возрастает, и необходимости длительного процесса классификации этих данных [2, 5].

Объединенные вместе данные проблемы создают предпосылки для перехода на использование более современной технологии анализа информации с использованием ИИ. Эти системы позволяют анализировать большие объемы данных, распознавать скрытые в них закономерности и эффективно принимать решения.

Методология исследования

Данные. В качестве основы был использован открытый датасет Home Credit Default Risk, включающий около 307 тысяч кредитных заявок и 122 признака для каждой из них. Данный набор сформирован финтех-компанией Home Credit, работающей на развивающихся рынках Восточной Европы и Азии, поэтому по своему социально-экономическому профилю он в определённой степени сопоставим с казахстанскими условиями.

В выборке представлены демографические характеристики (возраст, пол, семейное положение), финансовые показатели (доход, сумма кредита, аннуитетный платёж), данные кредитных бюро, а также информация о предыдущих займах [3].

Целевая переменная задаётся бинарно: значение 1 соответствует дефолту, 0 — его отсутствию. При этом доля дефолтов составляет около 8 %, что создаёт выраженный дисбаланс классов. На практике это означает, что без дополнительной настройки модель склонна «смещаться» в сторону большинства и предсказывать отсутствие дефолта почти для всех наблюдений, что формально даёт высокую accuracy, но не имеет практической ценности.

Валидация. Для корректного сравнения моделей использовалась стратифицированная пятикратная кросс-валидация, при которой сохраняется исходная пропорция классов в каждом разбиении. Основной метрикой качества выбрана ROC AUC, поскольку она отражает способность модели ранжировать заёмщиков по уровню риска независимо от выбранного порога.

Дополнительно рассматривались PR AUC (более чувствительная к редкому классу) и статистика Колмогорова–Смирнова (KS), широко применяемая в банковской практике для оценки разделяющей способности моделей.

Предобработка данных

Как правило, исходные кредитные данные требуют предварительной обработки: в них присутствуют пропуски, аномальные значения и технические заглушки.

Устранение аномалий. Например, в признаке DAYS_EMPLOYED встречалось значение 365243, что очевидно не соответствует реальному стажу и интерпретируется как специальный код отсутствующих данных. Такие значения были заменены на пропуски. Возраст, представленный в отрицательных днях, был преобразован в годы.

Финансовые переменные (доход, сумма кредита, аннуитет) характеризовались выраженной асимметрией распределения. Для её снижения применялось логарифмирование, что позволило сделать данные более устойчивыми для моделирования.

Производные признаки. Практика показывает, что абсолютные значения показателей информативны сами по себе лишь частично. Гораздо более показательны их соотношения. В связи с этим были сформированы дополнительные признаки:

$$PAYMENT\_RATE = AMT\_ANNUITY / AMT\_CREDIT \quad (1)$$

$$CREDIT\_INCOME\_RATIO = AMT\_CREDIT / AMT\_INCOME\_TOTAL \quad (2)$$

Эти показатели отражают уровень финансовой нагрузки: чем они выше, тем выше потенциальный риск дефолта.

Внешние рейтинги. Переменные EXT_SOURCE_1, EXT_SOURCE_2 и EXT_SOURCE_3 представляют собой агрегированные оценки из внешних источников. В рамках работы были рассчитаны их среднее и медианное значения, а также добавлены индикаторы наличия данных по каждому источнику, что позволило учитывать не только величину оценки, но и сам факт её доступности [3].

Очистка признакового пространства. Признаки с долей пропусков более 98 % были исключены. Категориальные переменные кодировались по-разному: для линейных моделей использовалось порядковое кодирование, для бустинговых алгоритмов - встроенные механизмы обработки категорий.

Числовые пропуски заменялись медианными значениями, категориальные - отдельной категорией «MISSING». В результате было сформировано устойчивое и сопоставимое признаковое пространство для всех моделей.

Модели машинного обучения

В рамках исследования были рассмотрены три группы алгоритмов: **логистическая регрессия**, **CatBoost** и **LightGBM**. Выбор именно этих методов обусловлен их распространённостью в задачах кредитного скоринга, а также различиями в подходах к моделированию — от линейных до ансамблевых.

Логистическая регрессия традиционно используется в скоринговых задачах и во многом остаётся базовым ориентиром. Модель предполагает, что вероятность дефолта определяется линейной комбинацией признаков, преобразованной с помощью сигмоидной функции:

$$P(y = 1 \mid x) = \sigma(\beta_0 + \Sigma_i \beta_i x_i), \text{ где } \sigma(z) = 1 / (1 + e^{-z}) \qquad (3)$$

Коэффициенты модели (β_i) интерпретируются достаточно прозрачно: они отражают вклад каждого признака в итоговую вероятность дефолта. Именно это свойство делает логистическую регрессию удобной для практического применения, особенно в условиях требований к объяснимости моделей [4].

В работе рассматривались два варианта реализации. В первом случае (LR1_select) предварительно проводился отбор признаков с использованием рекурсивного исключения и контроля мультиколлинеарности (VIF). Во втором варианте (LR2_regularized) применялась регуляризация:

$$\min_{\beta} [-\Sigma_i y_i \log(p_i) + (1-y_i) \log(1-p_i) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2] \qquad (4)$$

L1-регуляризация способствует обнулению менее значимых коэффициентов, тогда как L2 ограничивает их величину. На практике это позволило получить небольшой, но стабильный прирост качества модели.

CatBoost представляет собой алгоритм градиентного бустинга, ориентированный на работу с категориальными признаками. В отличие от классических подходов, он не требует предварительного кодирования категорий, что упрощает подготовку данных.

Модель формируется как сумма деревьев решений:

$$\hat{y} = \sigma(\Sigma_{t=1}^{T} \eta_t f_t(x)) \qquad (5)$$

Одной из ключевых особенностей является использование так называемого упорядоченного бустинга. При кодировании категориальных признаков алгоритм опирается только на часть данных, что позволяет снизить эффект утечки целевой переменной и уменьшить переобучение [6].

В рамках эксперимента были протестированы несколько конфигураций: базовая модель (CB1_quick), вариант со стохастической подвыборкой (CB5_subsample) и финальная настроенная версия (CB6_tuned). Настройка гиперпараметров дала пусть и незначительный, но устойчивый прирост качества.

Алгоритм **LightGBM** также относится к градиентному бустингу, однако отличается стратегией построения деревьев. В отличие от «уровневого» роста, здесь используется подход, при котором на каждом шаге расширяется наиболее перспективный лист.

Функция потерь задаётся следующим образом:

$$L = \Sigma_i\, l(y_i, \hat{y}_i) + \Omega(f_t),\ где\ \Omega(f_t) = \gamma T + \tfrac{1}{2}\lambda\|w\|^2 \qquad (6)$$

Такой подход позволяет ускорить обучение и, как правило, добиться более высокой точности, хотя при этом возрастает риск переобучения на небольших выборках [7].

Дополнительным преимуществом является использование гистограммного метода разбиения, что существенно снижает вычислительную сложность при работе с большими объёмами данных.

В процессе настройки подбирались ключевые параметры модели, включая число листьев (num_leaves), скорость обучения (learning_rate), долю используемых признаков (feature_fraction) и минимальное число наблюдений в листе (min_data_in_leaf). По итогам эксперимента именно LightGBM продемонстрировал наилучший результат.

Общие приёмы. С учётом выраженного дисбаланса классов (около 8 % дефолтов) использовалось взвешивание наблюдений, что позволило повысить чувствительность моделей к редкому классу.

Для бустинговых алгоритмов применялась ранняя остановка обучения, предотвращающая переобучение. Кроме того, выполнялась калибровка вероятностей: для логистической регрессии — методом Платта, для ансамблевых моделей — изотонической регрессией. Это обеспечило более корректную интерпретацию прогнозов в терминах вероятности дефолта [4, 5].

Результаты и обсуждение

Итоговые значения ROC AUC для всех моделей собраны в Таблице 1.

Таблица 1 - Сравнение моделей по метрике ROC AUC

Модель	Тип	AUC	Комментарий
LR1_select	Лог. регрессия (отбор признаков)	≈0,730	Базовая линейная модель
LR2_regularized	Лог. регрессия + L1/L2	≈0,750	Регуляризация улучшила обобщение
ANN_tuned	Нейронная сеть (MLP)	≈0,755	Стабильный, но невысокий результат
CB1_quick	CatBoost (по умолчанию)	≈0,765	Неплохой результат без настройки
CB5_subsample	CatBoost (подвыборка)	≈0,769	Подвыборка добавила устойчивости
CB6_tuned	CatBoost (настроенный)	≈0,773	Лучший из семейства CatBoost
LGB_tuned	**LightGBM (настроенный)**	**≈0,810**	**Лучшая модель в эксперименте**

Картина результатов в целом выглядит достаточно однозначной. Лидирующую позицию занимает LightGBM с AUC на уровне около 0,81, что примерно на восемь процентных пунктов выше базовой логистической регрессии. Для задач кредитного скоринга такая разница уже имеет практическое значение: модель заметно лучше ранжирует заёмщиков по уровню риска. В прикладном смысле это даёт банку больше гибкости - можно либо увеличить долю одобрений среди надёжных клиентов, либо сократить количество дефолтов при сохранении текущей политики [3, 7].

CatBoost демонстрирует второй по качеству результат. При этом разница между базовой конфигурацией (AUC ≈ 0,765) и настроенной версией (AUC ≈ 0,773) оказывается незначительной, но устойчивой. Это косвенно подтверждает, что подбор гиперпараметров даёт эффект, хотя и не радикальный. Вероятно, важную роль играет механизм

упорядоченного бустинга, позволяющий снизить переобучение на категориальных признаках, доля которых в данных достаточно велика [6].

Логистическая регрессия, как и ожидалось, показывает более скромные результаты. Тем не менее её нельзя считать неэффективной: значение AUC около 0,75 для регуляризованной модели остаётся приемлемым, особенно с учётом полной интерпретируемости. В задачах, где требуется прозрачность принятия решений, это может оказаться критически важным фактором [4].

Нейронная сеть (MLP) продемонстрировала результат на уровне AUC ≈ 0,755 - немного выше логистической регрессии, но заметно ниже бустинговых алгоритмов. Подобное поведение в целом ожидаемо: для табличных данных умеренного объёма ансамбли деревьев зачастую оказываются более эффективными, чем глубокие модели [2, 3].

Отдельного внимания заслуживает работа с дисбалансом классов. Без применения взвешивания все модели фактически сходились к тривиальному сценарию — предсказанию отсутствия дефолта для большинства наблюдений. Несмотря на формально высокую accuracy, такие результаты не имеют практической ценности. Использование взвешивания, а также ранней остановки позволило стабилизировать обучение. Дополнительная калибровка вероятностей сделала прогнозы более пригодными для расчёта ожидаемых кредитных потерь в рамках МСФО 9 [5].

Немаловажную роль сыграла и инженерия признаков. Производные показатели, такие как PAYMENT_RATE и CREDIT_INCOME_RATIO, стабильно входили в число наиболее значимых. Это ещё раз подтверждает, что относительные характеристики финансовой нагрузки оказываются информативнее абсолютных значений. Аналогичная ситуация наблюдается и с переменными EXT_SOURCE: их исключение приводило к заметному снижению качества модели.

С практической точки зрения важно учитывать и вычислительные затраты. LightGBM обрабатывает выборку объёмом более 300 тысяч записей за считанные секунды, что делает его удобным для использования в онлайн-сценариях. CatBoost требует несколько больше времени на обучение, однако это не является критичным ограничением. Логистическая регрессия остаётся наиболее быстрой, тогда как нейронная сеть демонстрирует наибольшие требования к ресурсам без соответствующего выигрыша в точности.

В целом LightGBM можно рассматривать как наиболее сбалансированное решение с точки зрения качества и скорости. В то же время в практических системах имеет смысл рассматривать ансамбли моделей — например, комбинацию LightGBM и CatBoost, — что позволяет повысить устойчивость прогнозов.

Заключение

Полученные результаты свидетельствуют о том, что переход от классических линейных моделей к алгоритмам градиентного бустинга действительно приносит пользу в виде повышенной точности модели. С наибольшей метрикой ROC AUC (≈0,810) оказался алгоритм LightGBM, значительно превзойдя базовую логистическую регрессию на восемь процентных пунктов. Также можно сказать, что алгоритм CatBoost показал хорошие результаты уже в условиях базового использования, эффективно обрабатывая категориальные признаки.

На практике это может оказать большое влияние на деятельность банковской сферы в Казахстане. Чем лучше работает модель скоринга, тем выше вероятность обнаружить проблемного заемщика. В конечном итоге, это поможет существенно снизить число дефолтов, причем даже небольшое повышение метрики скоринга способно дать большой эффект, принимая во внимание размеры существующего кредитного портфеля. Отдельным плюсом является возможность использования откалиброванных вероятностей дефолта в расчете резерва в соответствии с МСФО 9.

Необходимо учесть, что внедрение такой модели не потребует серьезных изменений в текущей системе, поскольку алгоритмы бустинга могут быть внесены в инфраструктуру банка в виде микросервисов, что позволит оперативно производить выдачу скорингов в режиме, близком к реальному времени, что актуально для цифровых банковских продуктов. Обеспечить контроль качества можно через стандартные метрики стабильности, такие как PSI.

С другой стороны, необходимо учесть и некоторые ограничения. Сначала нужно отметить, что использован набор данных с открытым доступом, который не полностью отражает специфику казахстанского рынка, а следовательно, результаты требуют валидации на реальных данных. Кроме того, в рамках этой работы не рассматривались вопросы алгоритмической справедливости, которые также представляют интерес с практической точки зрения.

Полученные результаты в целом соответствуют предварительным ожиданиям. LightGBM показал хорошие результаты в скоринге для оценки кредитоспособности человека.

Список использованной литературы:

1. Baglarbasi, E. (2025). Utilizing AI for Improved Credit Risk Assessment. Doctoral Dissertations and Theses, 61. Harrisburg University of Science and Technology. https://digitalcommons.harrisburgu.edu/dandt/61/

2. Liu, Y., Huang, F., Ma, L., Zeng, Q., & Shi, J. (2024). Credit scoring prediction leveraging interpretable ensemble learning. Journal of Forecasting. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3033

3. Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. Procedia Computer Science, 174, 158–163.

4. Qi, J., Yang, R., & Wang, P. (2021). Application of explainable machine learning based on Catboost in credit scoring. Journal of Physics: Conference Series, 1955(1), 012039. IOP Publishing.

5. Ponsam, J. G., Gracia, S. V. J. B., Geetha, G., et al. (2021). Credit Risk Analysis using LightGBM and a comparative study of popular algorithms. International Conference on Computing and Communications Technologies (ICCT). IEEE.

6. Yang, R., Wang, P., Li, L., & Yong, S. (2025). An explainable SSA-CatBoost machine learning model and application in corporate credit rating. Annals of Operations Research. https://link.springer.com/article/10.1007/s10479-025-06513-y

7. Zhou, L., Fujita, H., Ding, H., & Ma, R. (2021). Credit risk modeling on data with two timestamps in peer-to-peer lending by gradient boosting. Applied Soft Computing, 108, 107477.

8. Ehrhardt, A. (2022). Reject inference in credit scoring: a survey and practical guide. https://adimajo.github.io/rejectinference.html