УДК 004.415

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СОВРЕМЕННЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КРЕДИТНОГО СКОРИНГА

Сафина Г.Ф.

к. ф.-м. н, доцент,

ФГБОУ ВО «Уфимский университет науки и технологий», Нефтекамский филиал,

Нефтекамск, Россия

Хаматдинов А.И.

студент,

ФГБОУ ВО «Уфимский университет науки и технологий», Нефтекамский филиал,

Нефтекамск, Россия

Аннотация

В статье рассматриваются различные современные методы машинного обучения, применяемые для оценки кредитоспособности физических лиц. Анализируются такие методы машинного обучения как логистическая регрессия, ансамбли (Random Forest, градиентный бустинг) и нейронные сети. Отдельное внимание уделено альтернативным данным. Делается вывод о целесообразности выбора модели в зависимости от стратегии банка.

Ключевые слова: кредитный скоринг, машинное обучение, логистическая регрессия, ансамбли, случайный лес, градиентный бустинг, нейронные сети, альтернативные данные.

A COMPARATIVE ANALYSIS OF MODERN MACHINE LEARNING METHODS FOR CREDIT SCORING

Safina G.F.

PhD, Associate Professor,

Neftekamsk branch of the Ufa University of Science and Technology,
Neftekamsk, Russia
Khamatdinov A.I.
student,
Neftekamsk branch of the Ufa University of Science and Technology,
Neftekamsk, Russia

Abstract
This article examines various modern machine learning methods used to assess individual creditworthiness. Machine learning methods such as logistic regression, ensembles (Random Forest, gradient boosting), and neural networks are analyzed. Special attention is given to alternative data. A conclusion is drawn regarding the appropriateness of model selection depending on the bank's strategy.
Key words: Credit scoring, machine learning, logistic regression, ensembles, random forest, gradient boosting, neural networks, alternative data.

В последние годы высокими темпами растёт рынок розничного кредитования. Вместе с этим растут кредитные риски, которые на себя берут банки [1-3]. Традиционные методы экспертной оценки перестают быть надёжными, они становятся всё более субъективными, трудоёмкими и плохо масштабируемыми. Поэтому всё больше получают распространение так называемые автоматизированные скоринговые системы, основанные на методах машинного обучения (Machine Learning, ML) [4-7].

В рамках данной работы проводится сравнительный анализ основных ML – методов, используемых для прогнозирования дефолта заёмщиков, а также оценить возможность использования альтернативных данных. Полученные результаты могут быть полезны кредитным организациям при выборе подхода к построению скоринговых моделей.

Начнём с логистической регрессии. Логистическая регрессия является статистическим методом, который в отличие от линейной регрессии, позволяет оценивать вероятность наступления дефолта в интервале от 0 до 1 [1, 3]. Данная модель имеет вид:

$$\ln\frac{p}{1-p} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \ldots + \omega_n x_n, \quad (1)$$

где p – вероятность дефолта, x_i $(i = 1, 2, \ldots, n)$ – характеристики заёмщика, ϖ_i $(i = 0, 1, 2, \ldots, n-1)$ – весовые коэффициенты, определяемые на обучающей выборке методом максимального правдоподобия [1, 3].

Преобразуя (1), получаем сигмоидную функцию, которая отображает любую взвешенную сумму вероятностных признаков в интервале (0, 1):

$$p = \frac{1}{1 + e^{-(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \ldots + \omega_n x_n)}}. \quad (2)$$

при z→+∞ вероятность $P(Y = 1) \to 1$ (высокий риск),

при z→−∞ вероятность $P(Y = 1) \to 0$ (низкий риск).

Данная модель (2) обладает рядом преимуществ: высокая интерпретируемость (коэффициенты показывают направление и силу влияния каждого фактора), простое обучение, низкие вычислительные затраты. Но в тоже время обладает и рядом недостатков: высокая чувствительность к выбросам и мультиколлинеарности (это уязвимости, при которых математические алгоритмы могут начать давать неверные и нестабильные прогнозы из-за плохого качества данных) [4, 6].

Следующим методом машинного обучения будут деревья решений. Деревья решений – это иерархические структуры, состоящие из правил вида «если…, то…». Каждый узел такого дерева содержит проверку значения какого-то признака, а листья этого дерева итоговую классификацию [5, 7].

К достоинствам этого метода относятся: наглядность (легко объяснить клиенту причину или причины отказа), возможность работать с пропусками данных, нет потребности в масштабировании. К недостаткам относятся: склонность к переобучению, нестабильность работы – небольшие изменения в данных могут очень сильно изменить структуру дерева. В качестве примера такого дерева приведу рисунок 1.

Рисунок 1 – Дерево решений на примере кредитоспособности клиентов банка

Далее идут ансамблевые методы. Они объединяют множество базовых моделей (обычно деревья решений) для повышения точности и устойчивости.

Как пример, можем рассмотреть случайный лес (Random Forest). Данный вид ансамблевого метода строит множество деревьев на разных случайных подвыборках данных и случайных подмножествах признаков. Итоговое решение принимается голосованием.

В качестве преимуществ данного ансамбля можно выделить: устойчивость к переобучению и к различным выбросам, возможность оценить важность признаков. Недостатками являются: потеря наглядности отдельных деревьев, потребность в большей памяти и времени обучения.

Следующим известным ансамблем является градиентный бустинг (XGBoost, LightGBM, CatBoost). Он строит деревья последовательно, а каждое следующее дерево обучается предсказывать ошибки предыдущих.

К его преимуществам относятся: высокая точность прогнозов, гибкая настройка, встроенные механизмы регуляции. К недостаткам относятся: склонность к переобучению при неправильном подборе параметров, сложность интерпретации, большие требования к вычислительным ресурсам.

Перейдём к нейронным сетям. В последние годы данный ML-метод получил широкую известность. Нейронные сети имитируют работу биологических нейронов. Эти сети состоят из слоёв взаимосвязанных нейронов, способных к самообучению путём выявления сложных, нелинейных и иерархических паттернов в данных. Благодаря этому нейронные сети обладают максимальной точностью для самых сложных задач, но требуют огромное количество данных.

Естественно, у них есть как преимущества, так и недостатки. Преимущества: максимальная гибкость, способность работать с изображениями и текстами, автоматическое выделение признаков. Недостатки: требуют огромный объём данных, интерпретация крайне затруднена, нуждаются в мощных вычислительных ресурсах, очень чувствительны к настройке гиперпараметров.

Также помимо традиционных анкетных данных (возраст, стаж, доход и так далее) и кредитной истории, всё чаще встречается использование альтернативных данных. К ним относятся:

– цифровая активность – частота использования мобильного; приложения банка:

– данные из социальных сетей – анализ связей, подписок, публикуемого контента;

– поведенческий анализ – шаблоны расходов, регулярность; поступлений денежных средств, кому совершаются переводы;

– данные сотовых операторов – продолжительность звонков, частота смены номеров.

Использование таких данных позволяет оценивать клиентов с «тонкой» кредитной историей (молодёжь, мигранты) или с отсутствующей кредитной историей. Однако есть и существенные недостатки: нарушение конфиденциальности, возможность скрытной дискриминации, правовая неопределённость.

Проведённый анализ показывает, что современные методы машинного обучения предоставляют широкий анализ для автоматизации кредитного скоринга. Логистическая регрессия является, пожалуй, «золотым стандартом» там, где важна интерпретируемость. Ансамблевые методы дают более высокую точность, но требуют осторожности в настройке и объяснении решений. Нейронные сети превосходят их в точности, но при наличии очень больших объёмов данных и мощных вычислительных ресурсов. Альтернативные данные открывают новые возможности для оценки «сложных» заёмщиков, однако их использование должно строго соответствовать законодательству о персональных данных (Федеральный закон №152-ФЗ) и этическим нормам.

Библиографический список:

1. Алексеева Д. Г. Банковское кредитование: учебник для вузов / Д. Г. Алексеева, С. В. Пыхтин. – 2-е изд., перераб. и доп. – М.: Юрайт, 2025. – 132 с.

2. Кипкеева А. М. Управление экономическими рисками: учебник для вузов / А. М. Кипкеева, О. И. Алиев. – М.: Юрайт, 2026. – 137 с.

3. Эйтшгтон В.Н., Анохин С.А. Прогнозирование банкротства: основные методики и проблемы. – М.: ИНФРА-М, 2007. –205 с.

4. Воронов М. В. Системы искусственного интеллекта: учебник и практикум для вузов / М. В. Воронов, В. И. Пименов, И. А. Небаев. – 2-е изд., перераб. и доп. – Москва: Издательство Юрайт, 2026. – 268 с.

5. Кириллова Е.А., Сафина Г.Ф. Преимущества применения spring data jdbc для упрощения работы с базами данных / В сборнике: Актуальные вопросы современной науки и образования. Материалы Всероссийской научно-практической конференции, посвященной 25-летию Нефтекамского филиала УУНиТ. – Уфа, 2025. – С. 75-79.

6. Основы программирования на языке Python: учеб. пособие / под ред. А. Ю. Богачева. – Краснодар: РЭА, 2023. – 135 с.

7. Платонов, А. В. Машинное обучение: учебное пособие для вузов / А. В. Платонов. – 2-е изд. – М.: Юрайт, 2026. – 89 с.