УДК 336.71(045)

ПРОГНОЗИРОВАНИЕ ВЕРОЯТНОСТИ ДЕФОЛТА КОРПОРАТИВНЫХ ЗАЕМЩИКОВ*

Кузнецов Максим Дмитриевич,
студент
факультета прикладной математики и информационных технологий,
Финансовый университет,
Москва, Россия
makskuznetsov19@gmail.com

Аннотация. *В настоящее время банки очень активно стали развивать свои кредитные процессы и все больше внимания уделяют моделям, помогающим в принятии решений. В статье рассматриваются подходы к оценке кредитного рейтинга корпоративных заемщиков на основе решения задачи бинарной классификации. Сначала представлен классический подход на основе построения модели логистической регрессии, который применяется большинством банков в России. Он связан с тем, что от моделей для корпоративных заемщиков требуется полная интерпретируемость их результатов, чего можно добиться только в случае линейной зависимости вероятности дефолта от факторов, объясняющих само событие дефолта. Затем рассматривается возможность применения моделей машинного обучения к задаче классификации заемщиков. Это тип моделей, в которых интерпретируемость результатов намного ниже, но есть возможность воспроизводить нелинейные зависимости. В заключении представлено сравнение точности всех моделей.*

Ключевые слова: *эконометрическая модель; эконометрический подход; кредитные риски; МСФО 9; анализ данных; машинное обучение; кредитный скоринг; логистическая регрессия; кредитные рейтинги*

THE PREDICTION OF THE PROBABILITY OF DEFAULT FOR CORPORATE BORROWERS

Kuznetsov Maxim Dmitrievich,
student,
Faculty of Applied Mathematics and Information Technology,
Financial University,
Moscow, Russia
makskuznetsov19@gmail.com

Abstract. *Currently, banks are actively developing their credit processes and pay more attention to models that help them in decision-making. The article presents an approach to assessing the credit rating of corporate borrowers by solving a binary classification task. First, the author presented the classical approach, based on constructing a logistic regression model — approach that is used by most banks in Russia. It is*

Научный руководитель: ***Бывшев В.А.,*** *доктор технических наук, профессор, профессор Департамента анализа данных, принятия решений и финансовых технологий, Финансовый университет, Москва, Россия.*

* Статья победителя IX Международного научного студенческого конгресса «Цифровая экономика: новая парадигма развития».

because the models for corporate borrowers require full interpretability of their results. However, it can be achieved only in the case of a linear dependence of the probability of default on the factors that explain the event of default. Further, the author considered the possibility of applying machine learning models to the borrowers` classification task. It is a type of model where the interpretability of the results is much lower. However, it is possible to reproduce nonlinear dependencies. In conclusion, the author presented a comparison of the accuracy of all models.

Keywords: *econometric model; credit risks; IFRS 9; data analysis; machine learning; credit scoring; logistic regression; credit ratings*

В связи с внедрением в последнее время банками стандартов МСФО 9 вопрос расчета резервов стал как никогда актуален. Если до МСФО 9 банки рассчитывали свои резервы исходя из табличных коэффициентов Банка России, то сейчас реализуется подход на основе внутренних рейтингов. Это означает то, что банкам позволят рассчитывать величину своих резервов по рейтингам, которые они присваивают заемщикам на основе математических моделей. Большую часть резервов можно отнести на корпоративный сегмент, где идет работа с крупными заемщиками, которые являются индивидуальными предпринимателями, крупными компаниями и группами компаний. Исходя из этого, можно говорить о том, что доля этих заемщиков в портфеле банков велика и требуется так выстраивать кредитный процесс, чтобы вовремя замечать заемщиков, которые могут выйти в дефолт. Отсюда и возникает задача кредитного скоринга для корпоративных заемщиков.

Для того чтобы перейти на новую систему резервирования, требуется получить одобрение у Центрального банка. Одним из обязательных требований является валидация моделей, используемых для целей МСФО 9 у регулятора. Основными требованиями к таким моделям являются:

- высокая прогнозная точность;
- интерпретируемость результатов.

Если говорить об интерпретируемости результатов моделей, то в случае с корпоративными заемщиками она необходима, так как объемы кредитов являются большими и требуется полное понимание, что привело к выставлению рейтинга.

Модель логистической регрессии. Эта модель, широко используемая на практике, способна решать задачу классификации, ответом которой является вероятность дефолта заемщика (в терминах задачи классификации заемщиков на дефолтных и недефолтных) [1, с. 5]. Стоит отметить, что логистическая регрессия — классическая эконометрическая модель, которая хорошо зарекомендовала себя в практическом применении для многих задач. Такая модель позволяет классифицировать заемщиков, аппроксимируя линейные взаимосвязи между факторами и вероятностью дефолта заемщика. Отметим, что параметры модели оцениваются методом максимального правдоподобия [2, с. 191]. Первоначально оценивается регрессия на латентную переменную модели, а ответ регрессии шкалируется при помощи логистической функции, которая и переводит непрерывный ответ латентной переменной в вероятность принадлежности заемщика к классу 1. Эту модель Банк России готов проверять на адекватность и позволять использовать ответы моделей для целей расчета резервов.

Далее приведем практический опыт автора в применении логистической регрессии для оценки кредитного рейтинга корпоративных заемщиков. Описанная ниже модель и сам подход были разработаны лично автором в период его работы в банке в должности риск-аналитика департамента рисков и теперь используется банком для расчета резервов. На первом этапе исследования автором был проведен анализ имеющейся научной и практической литературы по теме построения моделей кредитного рейтинга и разработан подход к решению практической задачи. Следует отметить, что по данной тематике практически отсутствуют русскоязычные источники, банки опираются исключительно на иностранную литературу. На втором этапе были произведены сбор, обработка, анализ данных и построение модели, включая разработку приложения на языке R для доступа к данным в базах данных, работе с ними и построения моделей логистической регрессии.

Исходные данные состоят из 14 626 наблюдений по реальным заемщикам банка. Доля дефолтных наблюдений в выборке довольно мала

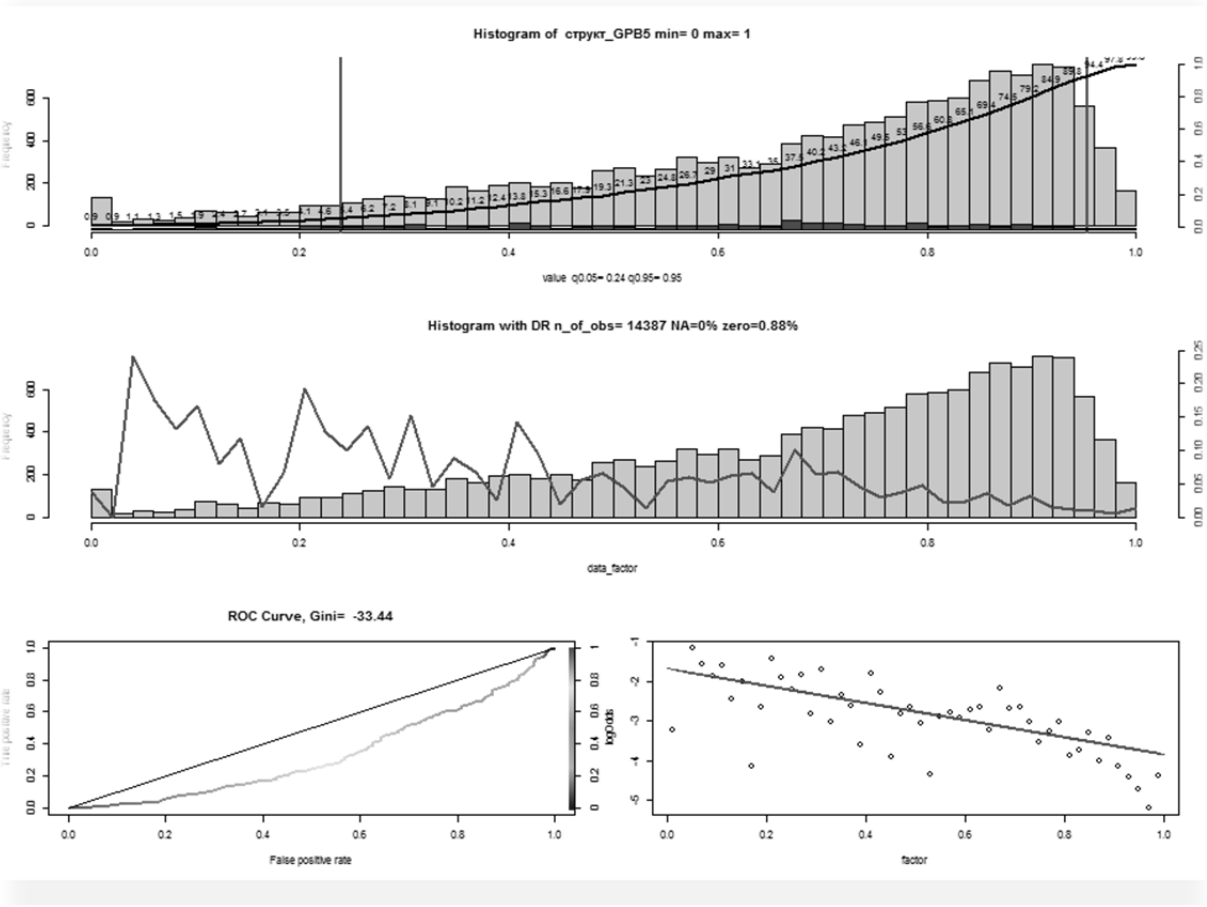

Рис. 1. **Однофакторный анализ**

Источник: составлено автором.

и составляет 672 наблюдения, т.е. 4,6% от общего числа наблюдений. Это довольно несбалансированный класс, так как дефолтных наблюдений очень мало, а основная задача банка состоит в отыскании именно тех заемщиков, которые не будут исполнять свои обязательства.

Сама модель состоит из двух основных блоков:

- финансовый блок — 70%;
- нефинансовый блок — 30%.

Финансовый блок строится на основе бухгалтерского баланса компании и отчета о движении денежных средств, исходная выборка представлена 64 объясняющими переменными. Стоит отметить, что все финансовые данные преобразуются, т.е. используются различные коэффициенты и отношения статей баланса. Для того чтобы все данные были одной размерности, проводится стандартизация данных и логистическое преобразование.

Нефинансовый блок представляет собой набор ординальных категориальных переменных, которые отражают макроэкономические показатели экономики и самой компании. В данной выборке мы имеем 25 экзогенных переменных.

Начальный этап исследования — проведение однофакторного анализа для финансовой модели. На данном этапе идет работа как со статистической информацией, так и с экспертами. Экспертно заранее определяется направление влияния фактора на вероятность дефолта заемщика. Данный факт обусловлен тем, что результат работы модели должен быть на 100% интерпретируем и непротиворечивым. Если, например, с ростом выручки компании растет вероятность дефолта, то такой фактор не может быть использован, так как он противоречит исходному смыслу. Многие значения факторов могут быть аномальными из-за различных ошибок ввода данных, выбросов. Для того чтобы сгладить выбросы, проводится винсоризация: если значение фактора превышает (меньше) квантиля заданного уровня, то этому значению присваивается значение выбранного

граничного квантиля. Винсоризация проводится как односторонняя, так и двусторонняя, что зависит от самого фактора и определяется при построении его гистограммы. При работе с данными также удаляются те факторы, которые имеют слишком много пропущенных значений, которые сильно коррелируют друг с другом (выбирается один наиболее значимый фактор среди всех коррелирующих).

Данный анализ является в большинстве своем визуальным и может быть представлен примером анализа одного из факторов (*рис. 1*). На рисунке верхний график показывает гистограмму распределения заемщиков, а также кумулятивную функцию распределения. Средняя гистограмма показывает, как с изменением значений фактора изменяется частота дефолтов заемщиков, например, на представленном примере можно заметить, что с ростом значения фактора частота дефолтов снижается. Нижние графики отражают силу разделяющей способности фактора.

На данном графике видно, что с ростом фактора вероятность дефолта падает. Фактор имеет неплохой показатель Gini, отражающий силу его разделяющей способности, аномалии сглажены. Линия на гистограмме в центре обозначает зависимость частоты дефолтов от значений фактора.

В результате однофакторного анализа удаляются факторы по следующим причинам:

- коррелированность с другими факторами;
- большое число пропущенных значений;
- направление влияния фактора отличается от экспертного мнения;
- оценка коэффициента в модели и показатель Gini нестабильны при изменении выборки.

Для нефинансового блока подход практически аналогичный, но вычисляются отдельно показатели WoE и IV, чтобы выставить штрафные баллы для значений фактора, так как такие факторы являются категориальными и, предполагается, ординальными категориальными переменными. Это означает наличие естественной упорядоченности значений категориальных переменных. Для категориальных переменных также проверяется корректность результатов расчета и интерпретация направлений влияния. Если оценки баллов какого-либо фактора не интерпретируемы, то они могут быть поправлены либо фактор полностью исключается из дальнейшего рассмотрения.

После отбора факторов можно приступать к многофакторному анализу. При формировании обучающего и тестового набора данных дефолтные и недефолтные наблюдения берутся в соотношении 1 к 3, что является эвристикой и было определено в ходе проведения экспериментов. Это требуется для того, чтобы в выборке не было избыточного количества хороших наблюдений, которые своим количеством размыли бы дефолтные наблюдения. Такая модель сильно ошибалась бы и не могла быть применена к решению практических задач [3, с. 431]. Также сохраняется «целостность данных», т.е. при разбиении на обучающее и тестовое множество ко всем дефолтным наблюдениям по заемщику добавляются все его недефолтные наблюдения. Коэффициенты модели логистической регрессии оцениваются методом максимального правдоподобия, но с учетом экспертного ограничения на знаки коэффициентов модели, что позволяет сразу на моменте обучения получать только те факторы, которые имеют правильный знак и согласуются с экономической интерпретацией. Разбиение данных и оценка модели производится определенное количество раз, например 1000.

В результате работы алгоритма выбирается наилучшая модель исходя из точности на тестовом и обучающем множестве: обучающее множество — 70% наблюдений, 30% отводится на тест.

На следующем этапе, после построения моделей финансовых и нефинансовых данных, требуется получить итоговую общую модель. Итоговая модель представляется в виде взвешенной суммы ответов двух моделей, что отражает формула

$$\text{Score} = 0,7 \times \text{Score}_{\text{fin}} + 0,3 \times \text{Score}_{\text{nofin}}$$

где: $\text{Score}_{\text{fin}}$ — вероятность дефолта по финансовой модели;

$\text{Score}_{\text{nofin}}$ — вероятность дефолта по нефинансовой модели.

Данная формула вводится для того, чтобы взвесить ответы двух моделей, получить их линейную комбинацию. Подбор взвешенных коэффициентов является также одним из этапов построения моделей оценки вероятности дефолта корпоративных заемщиков. Банки подходят к решению данного вопроса по-разному. Существует практика подбора весов эвристическими методами, когда аналитики больше доверяют результатам финансовых моделей и придают им больший

вес в линейной комбинации. В нашем случае коэффициенты подбирались математически таким образом, чтобы линейная комбинация давала наилучший результат на обучающем множестве (на тестовом множестве данные коэффициенты подбирать в целом неверно, это будет говорить о том, что модель «заглядывает в будущее», а значит, данный подход ошибочен). Есть также подход к построению общей модели, которая сразу включает в себя как финансовые экзогенные переменные, так и нефинансовые, и тогда не будет необходимости составления линейной комбинации ответов моделей. Но в российской практике такой подход редко используется в силу сложившейся практики построения моделей у банков, а также невысокой степени доверия к нефинансовым показателям как к значимым и существенным характеристикам кредитного риска.

Итак, представленный алгоритм позволил нам построить модель логистической регрессии для оценки вероятности дефолта корпоративных заемщиков: модель имеет точность Gini 54,81 или AUROC 77,41, что является неплохим результатом, и такая модель может быть применена для поставленных задач. Уместно отметить, что логистическая регрессия применяется по причине высокой степени интерпретации ее результатов, что крайне важно при работе с крупными клиентами, когда объемы кредитов очень велики.

Модели машинного обучения. Альтернативным подходом можно считать применение машинного обучения. Основное отличие такого подхода заключается в том, что указанные модели способны воспроизводить нелинейные взаимосвязи между экзогенными переменными и целевой переменной. Но данный факт ведет к тому, что результаты моделей становится сложно интерпретировать: если модель скажет, что вероятность дефолта заемщика равняется 0,99, то мы не сможем однозначно сказать, какие показатели привели к такому результату. Перед автором была поставлена задача изучить различные модели, основанные на машинном обучении, и разработать программный код на языке Python для построения этих моделей как продолжение исследования после построения основной модели, используемой банком.

Автором проанализированы результаты применения следующих моделей: XGBoost, LightGBM, Random Forest, нейронные сети, метод опорных векторов с радиально базисным ядром, логистическая регрессия, линейный метод опорных векторов. Уточним терминологию:

XGBoost — метод градиентного бустинга над деревьями решений. Данный алгоритм базируется на построении большого числа неглубоких деревьев, каждое из которых улучшает решение предыдущего. Ответ модели — сумма ответов всех деревьев. Данный алгоритм сильно склонен к переобучению, что требует большого внимания к этому факту при разработке.

LightGBM — метод градиентного бустинга над деревьями решений, который по своему алгоритму похож XGBoost. Он имеет небольшие отличия в части работы с непрерывными величинами для экономии памяти, а также алгоритм деления самого дерева настроен иначе, что позволяет также увеличить скорость обучения, хотя само количество деревьев в алгоритме обычно выше, чем у градиентного бустинга.

Random Forest — метод построения случайных деревьев на различных подвыборках. При помощи статистического бутстрэпа формируется большое число случайных подвыборок, размер которых совпадает с размером исходной выборки. Строятся глубокие деревья на каждой из подвыборок, после чего ответы деревьев усредняются для получения итогового значения. Случайные леса меньше склонны к переобучению [4, с. 410] в отличие от градиентного бустинга, а также в силу своих особенностей заметно снижают ошибку прогноза [5, с. 98].

Нейронные сети — модель машинного обучения, которая состоит из входного слоя нейронов (объясняющие переменные), скрытых слоев, выходного слоя (ответ модели) и активационных функций [6, с. 1116]. В самом простом виде без скрытых слоев нейронная сеть может представлять собой регрессионную модель либо модель логистической регрессии в зависимости от типа выбранной активационной функции. Нейронные сети способны воспроизводить очень сложные нелинейные взаимосвязи, но выбор правильной архитектуры таких моделей является нетривиальной задачей, требуя наличия опыта. Для работы нейронных сетей исходные данные часто приходится преобразовывать, так как они чувствительны к различным размерностям входных параметров [7, с. 155].

Заметим, что логистическая регрессия, описанная в первой части статьи, может считаться

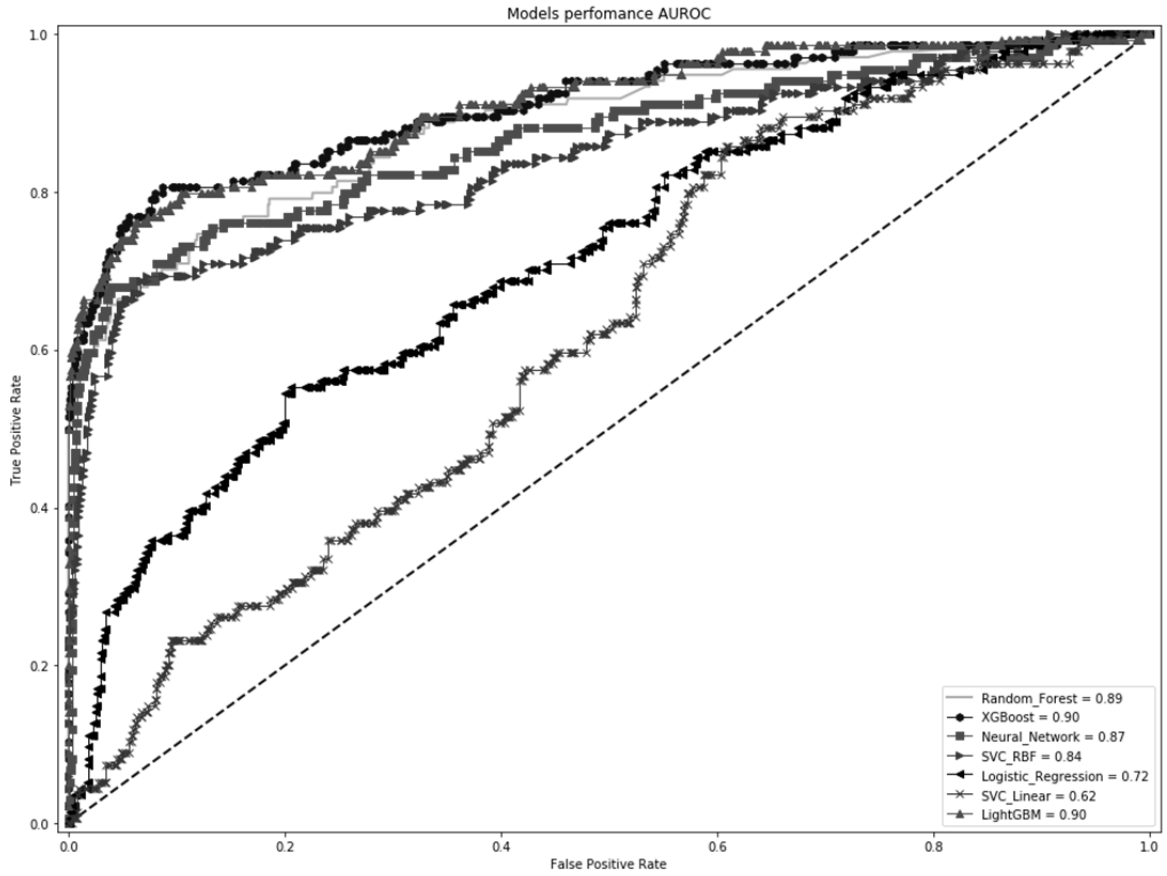

Рис. 2. **Сравнение точности моделей**

Источник: составлено автором.

не только эконометрической моделью, но также и моделью машинного обучения.

Линейный метод опорных векторов похож на метод логистической регрессии, но его суть заключается в проведении такой разделяющей поверхности, которая оставляет максимальный «зазор» между классами. Недостаток метода часто заключается в том, что его ответом является метка класса (0,1), что не всегда подходит для задач классификации [8, с. 743].

Последний метод — это метод опорных векторов с радиально-базисной функцией ядра. Часто в задачах бывает так, что классы не могут быть линейно разделены. Ядерные функции позволяют перейти в пространство большей размерности, где классы могут быть линейно разделимы. В случае с радиально-базисной функцией в исходном пространстве классы разделяются не гиперплоскостями, а сферами [9, с. 77].

Сами модели машинного обучения склонны к переобучению, т.е. склонны иногда просто «заучить» ответы, что приведет к очень низким точностям на данных, которые модель еще не видела в процессе обучения. Для борьбы с переобучением используют множество методов:

- выделение обучающего и тестового набора данных;
- кросс-валидация при обучении моделей [9, с. 175];
- подбор оптимальных гиперпараметров моделей.

Из исходного набора данных выделяется 80% данных на обучающую выборку и 20% для проверки качества моделей. Критерием точности будет метрика AUROC, которая имеет смысл вероятности того, что случайному дефолтному заемщику будет присвоено значение вероятности дефолта выше, чем случайному недефолтному.

На отобранных 80% данных проводится подбор оптимальных параметров каждой модели, где для каждой выбирается отдельный пул параметров. Подбор идет при помощи кросс-валидации с делением обучающей выборки на 3 блока. Модель обучается на каждой паре двух блоков и проверяет свою точность на оставшемся третьем. Общая точность — средняя точность на

каждом блоке. Таким образом, мы подбираем оптимальные параметры, контролируя переобучение модели.

Стоит отметить, что в нашем случае модели строились исключительно на финансовых данных и состояли только из одной модели без включения макроэкономических показателей.

Полученная модель тестируется на отложенной выборке. Результат на этом тесте и будет являться конечной точностью модели. На *рис. 2* представлен график получившихся результатов по разным моделям.

Как видно из графика, лучший результат показали модели XGBoost и LightGBM. Обе модели сразу классифицируют правильно более 60% дефолтных заемщиков с вероятностью дефолта 1, что является очень хорошим результатом. Видно, что модели машинного обучения являются хорошей альтернативой логистической регрессии в плане точности классификации заемщиков.

Сравнение точности моделей. Мы рассмотрели различные подходы к построению модели, позволяющей оценивать вероятность дефолта корпоративных заемщиков, что является актуальной задачей для большинства российских банков. Подходы к решению этой задачи различны: классический эконометрический подход на основе логистической регрессии и подход на основе машинного обучения. Машинное обучение в силу сложности своих моделей дает исследователям возможность существенно выиграть в точности, так как эти модели могут воспроизводить более сложные зависимости. Проблема такого метода заключается в том, что интерпретировать результаты становится очень сложно. В таких моделях можно говорить о том, какие переменные вообще имеют влияние и насколько сильное, но это относится ко всем наблюдениям в общем, а точечный анализ каждого заемщика этот подход нам провести не дает.

Подход на основе логистической регрессии, который в основном и применяется, хорошо зарекомендовал себя в решении данной задачи. Результаты отлично интерпретируются, наличие линейной зависимости также представляет собой ценное знание. Безусловно, представленная точность модели является только модельной. В банковской практике в помощь моделям используют различные ручные корректировки и предупреждающие сигналы. Основная их задача — снизить полученный рейтинг или даже перевести его в дефолтный при наличии оснований полагать, что у заемщика произойдет событие дефолта. Таким образом, мы констатируем, что при высокой точности машинного обучения банковская сфера не готова к его применению в задачах, связанных с корпоративным сегментом. Логистическая регрессия вполне способна решать необходимые банку задачи, удовлетворяя все требования регулятора. Однако сама идея применения различных методов, несомненно, должна применяться, поскольку это позволяет лучше понять данные и в итоге выбрать оптимальную модель.

Список источников

1. *The Basel II Risk Parameters: Estimation, Validation and Stress Testing with Application to Loan Risk Management*. Engelmann B., Rauhmeier R., eds. 2 ed. Heidelberg, Dordrecht, London, New York: Springer; 2011. 419 p.
2. Носко В.П. *Эконометрика.* Кн. 2. Ч. 3, 4. Учебник. М.: Дело; 2011. 576 с.
3. Leong C.K. Credit Risk Scoring with Bayesian Network Models. *Computational Economics*. 2016;47(3):423–446.
4. Barboza F., Kimura H., Altman E. Machine Learning Models and Bankruptcy Prediction. *Expert Systems with Application*. 2017;83(C):405–417.
5. Ala'raj M., Abbod M.F. Classifiers consensus system approach for credit scoring. *Knowledge-Based System*. 2016;104:89–105.
6. Tsai C.-F., Hung C. Modeling credit scoring using neural network ensembles. *Kybernetes*. 2014;43(7):1114–1123.
7. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб.: Питер; 2018. 480 с.
8. Harris T. Credit scoring using the clustered support vector machine. *Expert systems with Application*. 2015.42(2):741–750.
9. Raschka S. *Python Machine Learning*. Birmingham, Mumbai: Packt Publishing; 2015. 454 p.