DOI: 10.17323/1998-0663.2019.3.52.66

Сравнительный анализ методов прогнозирования банкротств российских строительных компаний

А.М. Карминский
E-mail: karminsky@mail.ru

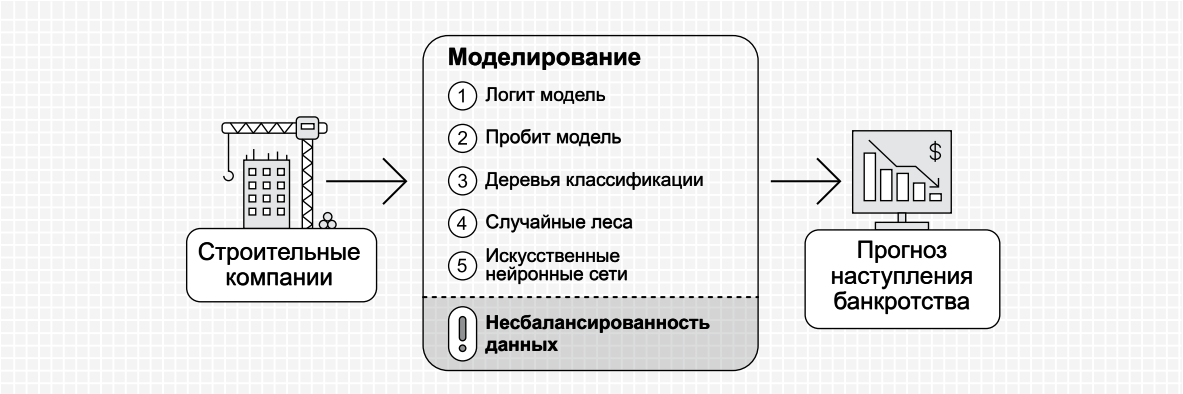
Р.Н. Бурехин
E-mail: romanvia93@yandex.ru

Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

Аннотация

Работа посвящена сравнению способности различных моделей предсказывать банкротство компаний строительной отрасли на горизонте в один год. Рассмотрены такие алгоритмы, как логит- и пробит-модели, деревья классификации, случайные леса, искусственные нейронные сети. Особое внимание уделено особенностям построения моделей машинного обучения, влиянию несбалансированности данных на предиктивную способность моделей, анализу способов борьбы с несбалансированностью данных, анализу влияния нефинансовых факторов на предиктивную способность моделей. В работе использованы нефинансовые и финансовые показатели, рассчитанные на основе публичной финансовой отчетности строительных компаний за период с 2011 по 2017 годы. Сделан вывод, что рассмотренные алгоритмы показывают приемлемое качество для использования в задачах прогнозирования банкротств. В качестве метрики качества моделей использовался коэффициент Джини или AUC (площадь под ROC-кривой). Выявлено, что искусственные нейронные сети превосходят другие методы, в то время как модели логистической регрессии в сочетании с дискретизацией вплотную следуют за ними. Обнаружено, что эффективность способа преодоления несбалансированности данных зависит от типа используемых моделей. В то же время значимого влияния несбалансированности обучающего множества на предиктивную способность модели не выявлено. Существенное влияние нефинансовых показателей на вероятность банкротства также не подтвердилось.

Графическая аннотация



Ключевые слова: банкротство; строительный сектор; несбалансированность данных; модели машинного обучения; параметрические модели предсказания банкротства.

Цитирование: Карминский А.М., Бурехин Р.Н. Сравнительный анализ методов прогнозирования банкротств российских строительных компаний // Бизнес-информатика. 2019. Т. 13. № 3. С. 52–66.
DOI: 10.17323/1998-0663.2019.3.52.66

Введение

В рыночной экономике прогнозирование финансовой несостоятельности является важной задачей для любой компании. Для достижения этой цели используются разные методы оценки кредитных рисков, целью которых является заблаговременное и эффективное прогнозирование наступления неблагоприятной ситуации на предприятии. Обычно эти методы представляют собой параметрические модели, характеризующиеся относительно несложным математическим аппаратом и простой качественной интерпретацией. Данные методы достаточно статичны, не учитывают тонких экономических или поведенческих факторов, а прогностическая способность моделей снижается при нелинейном характере связей между показателями.

Рыночные модели (структурные модели и модели сокращенной формы) часто оказываются слишком сложными или зависимыми от рынка. Для их применения необходим доступ к большому массиву данных (рыночной стоимости акционерного капитала, долговых обязательств, спрэдов доходности облигаций и т.д.). Несмотря на широкое применение рыночных моделей западными компаниями, на российском рынке их использование затруднено из-за небольшого количества котирующихся ценных бумаг. Для проведения эффективной кредитной политики новые методы должны быть гибкими и адаптируемыми к изменяющимся реалиям рыночной экономики. Поэтому в настоящее время имеется интерес к моделям на основе алгоритмов машинного обучения, включая деревья классификации, случайные леса, градиентный бустинг, искусственные нейронные сети и т.д.

Есть ряд общих проблем, связанных с прогнозом банкротств компаний. Во-первых, экономические показатели, описывающие состояние фирмы, отличаются в различных исследованиях, а их объединение в наиболее эффективную модель вызывает дополнительные сложности. Во-вторых, существует проблема несбалансированности данных, поскольку платежеспособных компаний больше, чем обанкротившихся. Как следствие, обученная модель имеет тенденцию классифицировать компании как надежные, хотя у них могут быть признаки финансовой несостоятельности. В-третьих, само понятие «банкротство» может интерпретироваться по-разному, поэтому в данную категорию могут попадать разные компании. В данной работе в категорию банкротов попали компании, в отношении которых начата процедура легального банкротства, а также компании, ликвидировавшиеся добровольно.

Несмотря на важность задачи прогнозирования банкротств с помощью более продвинутых методов, отечественных работ в данной области не так много, а работы, посвященные прогнозированию отзывов банковских лицензий, скорее являются исключением [1, 2]. Особенностью данной работы является сравнение регрессионных моделей и моделей на основе методов машинного обучения в задачах прогнозирования банкротств компаний на базе одной отрасли. Важное внимание уделяется специфике построения моделей машинного обучения, влиянию несбалансированности данных, а также нефинансовых показателей на предиктивную способность моделей.

Строительная отрасль является связующим звеном между другими отраслями, что обуславливает ее важность в национальной экономике. На сегодняшний день в России насчитывается более двухсот семидесяти тысяч компаний, выполняющих те или иные строительные работы (проектирование, инженерные расчеты, строительство и т.п.). Их количество, а также высокий уровень дефолтов в данном секторе приводит к сложности выбора подходящего партнера. Данная отрасль является одной из наиболее сильно пострадавших от кризиса. В частности, объем работ в сопоставимых ценах не перестает падать с 2014 года, а по итогам 2017 года строительство оказалось отраслью с одной из самых высоких долей безнадежных долгов. Кредитование строительного сектора представляет собой значительную часть деятельности российского банковского бизнеса, поэтому увеличение числа неплатежеспособных строительных компаний может стать при-

чиной нестабильности в банковском секторе. Более того, национальные и международные регуляторные требования (рекомендации Базельского комитета) вынуждают использовать продвинутый подход на основе внутренних рейтингов для количественной оценки рисков с целью снижения нагрузки на капитал. Поэтому проблема прогнозирования будущего состояния строительных компаний является актуальной, а новые инструменты прогнозирования банкротств востребованными.

Данная работа отвечает на вопрос, могут ли модели на основе методов машинного обучения быть достойной альтернативой регрессионным моделям при применении в области прогнозирования банкротств компаний нефинансового сектора, на примере строительной отрасли. Делается вывод, что все рассмотренные модели способны предсказать банкротство в ближайшие 12 месяцев, при этом нейронные сети превосходят другие методы в задаче идентификации неплатежеспособных компаний, а модели логистической регрессии в сочетании с дискретизацией вплотную следуют за ними. Негативного влияния несбалансированности обучающего множества на предиктивную способность модели обнаружено не было[1].

1. Модели прогнозирования финансовой несостоятельности

Регрессионные модели (логит- и пробит-модели) распространены в задачах идентификации платежеспособных и неплатежеспособных заемщиков [3]. Их преимущество заключается в отсутствии жестких ограничений функционирования, легкости интерпретации, простоте расчетов. Важным недостатком данных моделей является снижение прогностической способности при нелинейном характере связей между показателями, в то время как алгоритмы машинного обучения менее чувствительны к данным проблемам. Существует множество работ, доказывающих возможность использования продвинутых методов прогнозирования несостоятельности компаний [4–7].

Авторы работы [8] были одними из первых, кто использовал деревья классификации для прогнозирования банкротств компаний. Они обнаружили, что их деревья классификации превосходят дискриминантный анализ. Также было отмечено, что с усложнением модели (включением большего числа факторов) происходило ухудшение ее точности, что связано с переобучением. Однако этот успех не стал причиной распространенного использования деревьев решений в рассматриваемой области. В дальнейшем в большинстве работ имеет место сравнение эффективности деревьев решений с другими алгоритмами. Алгоритм случайных лесов был представлен в работе [9] и применен во многих областях: от маркетинга (прогнозирование лояльности клиента к бренду) и уголовной сферы (прогнозирование убийств или рецидивов среди условно-досрочно освобожденных), до кредитного скоринга. Опираясь на данные финансовой отчетности, авторы работы [4] успешно используют модели случайного леса при прогнозировании дефолтов компаний из семи европейских государств (Финляндии, Франции, Германии, Италии, Португалии, Испании и Великобритании). В 1990 году авторы работы [10] были одними из первых, кто использовал нейронную сеть при прогнозировании банкротств. Нейронная сеть строилась с несколькими скрытыми слоями и применением в качестве входных данных финансовых коэффициентов, используемых в модели Альтмана. При этом доля правильно классифицированных компаний составляла около 80%.

Данные алгоритмы часто показывают более высокую эффективность, несмотря на то, что они характеризуются значительными временными и физическими затратами. Более того, в настоящее время наблюдается тенденция, при которой алгоритмы, базирующиеся на одном методе, теряют популярность, в то время как ансамблевые или гибридные модели становятся все более популярными и демонстрируют более высокую эффективность [11].

С 1970-х годов финансовые коэффициенты, полученные на основе финансовой отчетности, являются важным источником построения моделей прогнозирования дефолтов. Однако модели, основанные на бухгалтерской информации, критикуются из-за отражения исторического характера информации, использующейся в качестве исходных данных и не учитывающей волатильность стоимости компании в течение анализируемого периода. При этом сторонники данного подхода утверждают, что неэффективность рынков капитала может приводить к более существенным ошибкам при прогнозировании кредитных рисков. В работе [12] сравниваются модели оценки кредитного риска на

[1] Предварительные результаты исследования представлены в выпускной квалификационной работе Р.Н. Бурехина, выполненной на факультете экономических наук НИУ ВШЭ в 2018 г.

основе бухгалтерской и рыночной информации. Авторы приходят к выводу, что значимых отличий в предсказательной способности у рассмотренных подходов нет, при этом эти типы данных являются взаимодополняющими, а комплексная модель показывает наилучший результат.

Можно сделать вывод, что рыночная информация может быть значимым фактором при прогнозировании несостоятельности компаний. Однако ввиду того, что большинство рассмотренных компаний не имеют выхода на фондовый рынок, финансовая отчетность становится единственным доступным источником информации, а применение рыночных моделей становится невозможным.

2. Описание используемых данных

Основным источником данных в данной работе является система СПАРК (агентство «Интерфакс»). Также при сборе информации о дефолте компаний была использована база данных «Единый федеральный реестр сведений о банкротстве». В исследовании к строительным компаниям были отнесены следующие компании (классификация в СПАРК):

- ✦ строительство зданий;
- ✦ строительство инженерных сооружений;
- ✦ специализированные строительные работы (разработка и снос зданий, подготовка строительного участка, отделочные строительные работы).

Важным вопросом является определение неплатежеспособной компании. В соответствии с Федеральным законом от 26.10.2002 № 127-ФЗ (редакция от 29.12.2017) «О несостоятельности (банкротстве)» [13] признаком банкротства юридического лица считается ситуация, когда требования кредиторов по денежным обязательствам не исполнены им в течение трех месяцев с даты, когда они должны были быть исполнены. В исследовательских работах широко распространены следующие определения банкротства: компания не способна оплатить проценты по долгу или часть основного долга, организация находится под наблюдением (процедура, при которой анализируется материальное положение и платежеспособность должника, а также его возможность погасить долговые обязательства), фирма не проявляет активности в течение длительного периода времени, компания находится в состоянии ликвидации. В нашей работе в категорию банкротов попали компании, в отношении которых запущена процедура легального банкротства, а также компании, ликвидировавшиеся добровольно. Подобная классификация приведена в работах [4, 14]. Отмечается, что данные компании отличаются критическим финансовым положением и часто не способны исполнять свои обязательства.

На основе финансовой отчетности были рассчитаны значения четырнадцати коэффициентов, отражающих хозяйственную деятельность предприятия. При этом была предложена следующая классификация финансовых показателей: рентабельность, ликвидность, деловая активность, финансовая устойчивость. Подобная классификация приведена в работе [3]. Также в модель включены нефинансовые факторы, отражающие размер и возраст компании. Описание переменных приведено в *таблице 1.*

Для анализа дефолтных событий был выбран временной диапазон 2011–2017 гг. Временной горизонт был разит на два блока: обучающую выборку (период 2011–2015 гг.) и тестовую выборку (период 2016–2017 гг.). На следующем этапе была проведена следующая процедура отбора:

1) удаление наблюдений с пропущенными данными (например, по которым отсутствует информация о величине активов и выручки) или заполнение пропусков в данных, где это возможно;

2) удаление наблюдений с очевидными ошибками (например, где размер активов или размер дебиторской задолженности отрицательный);

3) выявление и удаление наблюдений-выбросов, поскольку их наличие приводит к смещенности результатов. Основной алгоритм, использованный для данной процедуры, – правило трех сигм.

В результате в финальную выборку попала 3981 организация, 390 из которых объявили дефолт. В обучающую выборку (период с 2011 по 2015 гг.) попало 3300 строительных компаний, из которых 325 допустили дефолт. В тестовую выборку (период с 2016 по 2017 гг.) попала 681 компания, из которых 65 допустили дефолт. На *рисунке 1* отражено общее количество компаний и количество компаний, обанкротившихся в следующий отчетный год, по годам.

Исследовательский набор данных является несбалансированном (всего 9,8% компаний объявили дефолт). Поэтому при построении моделей были использованы две техники работы с несбалансированными данными: андерсемплинг (undersampling) и оверсемплинг (oversampling).

Андерсемплинг подразумевает использование входных данных, содержащих все неплатежеспособные компании, и случайную выборку платежеспособных

Таблица 1.

Описание используемых переменных

Группа	Переменные	Описание переменной
Зависимая переменная	Банкротство	1 – если произошел дефолт в следующем отчетном периоде; 0 – иначе
Рентабельность	Рентабельность активов	Отношение чистой прибыли к активам
	Рентабельность собственного капитала	Отношение чистой прибыли к собственному капиталу
	Рентабельность продаж	Отношение чистой прибыли к выручке
	Операционная маржа	Отношение операционной прибыли к выручке
Ликвидность	Коэффициент текущей ликвидности	Отношение оборотных активов к краткосрочным обязательствам
	Коэффициент быстрой ликвидности	Отношение суммы дебиторской задолженности, финансовых вложений и денежных средств к краткосрочным обязательствам
	Коэффициент маневренности собственного капитала	Отношение разности между собственным капиталом и внеоборотными активами к собственному капиталу
Деловая активность	Коэффициент оборачиваемости дебиторской задолженности	Отношение выручки к дебиторской задолженности
	Коэффициент оборачиваемости кредиторской задолженности	Отношение себестоимости продаж к кредиторской задолженности
	Коэффициент оборачиваемости активов	Отношение выручки к активам
	Доля внеоборотных активов	Отношение внеоборотных активов к активам
Финансовая устойчивость	Коэффициент автономии	Отношение собственного капитала к активам
	Доля нераспределенной прибыли в выручке	Отношение нераспределенной прибыли к выручке
	Коэффициент покрытия процентов	Отношение суммы прибыли до налогообложения и процентов к уплате к процентам к уплате
Размер компании	Логарифм активов компании	Логарифм активов
Возраст	Возраст	

Рис. 1. Количество банкротств по годам

компаний. В результате соотношение неплатежеспособных и платежеспособных компаний увеличивается. Также при построении такого рода зависимостей рекомендуется проводить данный эксперимент несколько раз для получения состоятельных результатов (в данном исследовании делается допущение о получении состоятельного результата после одного эксперимента). Оверсемплинг предполагает использование входных данных, содержащих все платежеспособные компании, и «клонирование» неплатежеспособных компаний до тех пор, пока их число не приблизится к числу платежеспособных компаний. Поиск оптимальной доли миноритарного класса на обучающем множестве также является предметом исследования в данной работе.

Для поиска оптимального значения доли неплатежеспособных компаний в андерсемплинге использовалась концепция кросс-валидации. При

использовании оверсемплинга использовать данный подход не рекомендуется, так как фактически в данном случае происходит клонирование информации, которая используется как при обучении, так и при тестировании модели (что приводит к переобучению). Для реализации оверсемплинга обучающее множество было разделено на два подмножества. Первое (информация о дефолтах компаний за 2014 год) использовалось для тестирования модели, второе (оставшиеся периоды с 2011 по 2015 гг.) – для построения модели без кросс-валидации. Доля неплатежеспособных компаний, при которой модель имеет наибольшее значение Джини на обучающем множестве, использовалась для сравнения моделей на тестовом множестве (2016–2017 гг.)

3. Описание моделей

В работе использовались два параметрических алгоритма построения моделей бинарного выбора: логит- и пробит-модели с поправками на дискретизацию (с использованием WOE) и без. Данные модели сравниваются с алгоритмами, основанными на методах машинного обучения (деревья классификации, случайные леса, искусственные нейронные сети), описание которых приведено в последующих разделах.

Традиционно выделяют следующие метрики качества моделей: точность (accuracy), чувствительность (sensitivity), специфичность (specificity), площадь под ROC-кривой, коэффициент Джини, F-метрика. Использование этих метрик зависит от цели анализа.

В данном исследовании каждая из рассмотренных моделей на выходе имеет область значений от 0 до 1, поэтому необходимо определить порог отсечения. Присвоение порога отсечения зависит от предпочтений аналитика относительно ошибок первого и второго рода, что приводит к трудностям при сравнении различных моделей. Поэтому в данной работе оценка предсказательной силы производится с помощью ROC-анализа, показателя AUC (площади под ROC-кривой) или коэффициента Джини. Преимущество данных метрик заключается в отсутствии необходимости определять порог отсечения и возможности сравнивать качество моделей независимо от целей аналитика. Расчет коэффициента Джини проводился следующим образом:

$$Gini = (AUC - 0{,}5) \cdot 2 \cdot 100\%, \qquad (1)$$

где AUC – площадь под ROC-кривой.

Визуальный анализ эффективности проводился с помощью ROC-кривой. Чем больше изгиб ROC-кривой, тем выше качество модели, при этом диагональная линия соответствует полной неразличимости двух классов. Соответственно, чем выше значение площади под ROC-кривой, тем лучше разделяющая сила модели. Анализ ROC-кривой позволяет пользователю выбрать необходимое для анализа соотношение между чувствительностью и специфичностью. Пример построения ROC-кривой для одной переменной представлен на *рисунке 2*.

В работе в качестве инструмента расчетов для эконометрического анализа и отражения статистических выводов был использован язык программирования R, представляющий собой свободную программную среду с открытым исходным кодом.

3.1 Модели бинарного выбора

В работе использовались два алгоритма построения моделей бинарного выбора (логит и пробит): без поправок на дискретизацию и с поправками. Приведем описание общего алгоритма (с поправками на дискретизацию) построения моделей бинарного выбора.

Этап 1. Приведение факторов к дискретной форме. В процессе решения исследовательского вопроса большинство авторов сталкиваются с проблемой выбросов. Эта проблема не является исключением и для данной работы. Традиционный подход для ее решения – исключение подобных наблюдений. Однако субъективность определения выбросов и уменьшение выборки являются существенными недостатками данного подхода. В работе используется

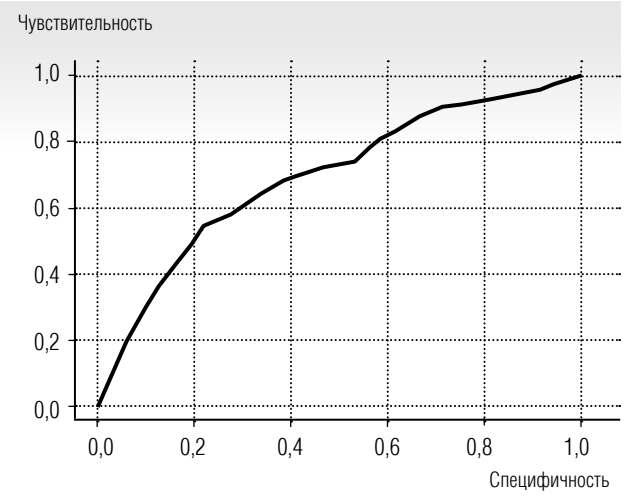


Рис. 2. ROC–кривая фактора ROA

переход от дискретной к непрерывной форме, что приводит к повышению сравниваемости факторов между собой и единству подходов к оценке значимости факторов. Нами проведена процедура квантильной дискретизации – замена исходных значений факторов на дискретные значения на основе группировки по квантилям. Суть данного подхода состоит в следующем:

а) значения переменных упорядочиваются по возрастанию;

б) значения каждого показателя делятся на 10 групп (используются децили);

в) значения в каждой группе заменяются на баллы от 1 до 10 (группа с наименьшими значениями получает 1 балл, а группа с наибольшими значениями – 10 баллов).

Этап 2. Преобразование факторов. Для преобразования факторов используется подход WOE (Weight of Evidence). Показатель WOE характеризует степень отклонения уровня дефолтов в данной группе от среднего значения в выборке. Для каждого фактора и для каждой группы внутри факторов необходимо вычислить число компаний в дефолте и число компаний не в дефолте. Показатель WOE для группы *i* определенной переменной рассчитывается следующим образом:

$$WOE_i = \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right), \qquad (2)$$

где $d_i^{(1)}$ – доля компаний не в дефолте, принадлежащих группе *i*, в общем числе компаний не в дефолте; $i = 1, 2, ..., k$; k – число категорий переменной;

$d_i^{(2)}$ – доля компаний в дефолте, принадлежащих группе *i*, в общем числе компаний в дефолте; $i = 1, 2, ..., k$; k – число категорий переменной.

С целью повышения линейности переменных и повышения точности модели все объясняющие переменные заменены на WOE, что является распространенным приемом в кредитном скоринге [15].

Этап 3. Оценка предсказательной силы факторов. После того как все значения преобразованы в WOE, необходимо оценить важность каждого фактора. В работе использованы два алгоритма оценки значимости факторов: информационная ценность (Information Value, *IV*) и ROC-анализ. Расчет значения информационной ценности (*IV*) выполнен по следующей формуле:

$$IV = \sum_{i=1}^{k}\left(d_i^{(1)} - d_i^{(2)}\right)\cdot WOE_i, \qquad (3)$$

где k – число категорий независимой переменной (у каждого фактора их десять), остальные обозначения из формулы (2).

Формула (3), отражающая расчет *IV*, базируется на суммировании WOE_j, скорректированного на разницу $\left(d_i^{(1)} - d_i^{(2)}\right)$. Основная цель данных расчетов – выявить некоторый индикатор, отражающий способность переменной кластеризовать какой-то признак. Если данный показатель выше 0,02, то фактор следует использовать при моделировании [15].

В исследовании применены следующие критерии отбора факторов в итоговую модель:

✦ приемлемое качество модели в соответствии с критерием «информационная ценность» ($IV > 0{,}02$);

✦ коэффициент Джини в однофакторной модели должен быть больше 5%;

✦ экономическое обоснование фактора.

Этап 4. Анализ корреляций. При построении многофакторной модели необходимо исключить факторы с высокими коэффициентами корреляции. Анализ корреляции позволяет избежать мультиколлинеарности. Мультиколлинеарность приводит к неустойчивости модели и повышает стандартные отклонения оценок факторов. О наличии мультиколлинеарности свидетельствуют высокие значения парных коэффициентов корреляции между факторами переменных. Критерий для определения высокой корреляции может варьироваться, для экономических данных порог обычно устанавливается на уровне 0,30–0,50. Критерием высокой корреляции в данной модели является коэффициент корреляции больше 0,5.

Этап 5. Многофакторный анализ. Моделирование вероятности некредитоспособности заемщика осуществлялось следующим образом:

$$\mathrm{P}\left(Y_i = 1 | x_1, \ldots, x_n\right) = F\left(a_0 + a_1 x_1 + \ldots + a_n x_n\right) = F\left(a_0 + \mathbf{x}'a\right). \qquad (4)$$

При этом в случае логит-модели F (*) представляет функцию логистического распределения:

$$F\left(a_0 + \mathbf{x}'a\right) = \Lambda\left(a_0 + \mathbf{x}'a\right) = \frac{e^{a_0 + \mathbf{x}'a}}{1 + e^{a_0 + \mathbf{x}'a}}. \qquad (5)$$

В случае пробит-модели F (*) представляет собой функцию нормального распределения:

$$F(a_0 + \mathbf{x}'a) = \Phi(a_0 + \mathbf{x}'a) = \int_{-\infty}^{a_0 + \mathbf{x}'a} \varphi(v)dv, \qquad (6)$$

где $\varphi(v) = \frac{1}{\sqrt{2\pi}} e^{\left(-v^2/2\right)}$.

Расчет коэффициентов осуществляется методом максимального правдоподобия, который максимизирует вероятность совместной реализации событий (платежеспособность и неплатежеспособность). Стандартные ошибки коэффициентов оценивались с поправкой Ньюи–Уэста на гетероскедастичность и автокорреляцию первого порядка.

Этап 6. Спецификация модели. Для выбора оптимальной комбинации факторов был использован метод обратного перебора (Backward Selection) – последовательное исключение факторов (т.е. из модели, в которую включены все факторы, отобранные при однофакторном анализе, последовательно исключаются незначимые переменные). При этом уровень статистической значимости тестируется с помощью *p*-value, рассчитанного по результатам логистической регрессии. В результате были отобраны факторы с *p*-value меньше 10%.

Этап 7. Валидация модели. Выбор лучшей модели. Проблема переобучения требует проводить процедуру валидации модели. Данная проблема проявляется в том, что «обученная» модель имеет хорошие результаты на обучающей выборке, но не дает точных прогнозов для тестовой выборки. Для решения данной проблемы было использовано два подхода.

Первый подход – «алгоритм перемешивания», идея которого сводится к следующему:

1. Случайным образом выбирается 80% компаний из обучающей выборки;

2. Оцениваются коэффициенты модели;

3. Оценивается, сохраняются ли при этом знаки при полученных коэффициентах, а также остается ли рассмотренный фактор значимым;

4. Шаги 1–3 повторяются 1000 раз, проверяется стабильность знаков.

Опираясь на полученные результаты, можно сделать вывод, являются ли знаки коэффициентов при всех переменных стабильными, и как знак коэффициента зависит от исходной выборки.

Второй подход – ROC-анализ. Анализ значений *AUC* и Джини на тестовом множестве помогает сделать вывод о качестве полученных моделей.

3.2. Модели на основе методов машинного обучения

Логистический анализ и пробит-анализ являются традиционными популярными инструментами прогнозирования банкротств, но имеют ряд недостатков, связанных с низкой прогностической силой и наличием ограничений по использованию. Поэтому на сегодняшний день получили распространение алгоритмы машинного обучения.

Деревья классификации. На сегодняшний день деревья классификации являются фундаментом для построения более сложных алгоритмов машинного обучения, таких как случайные леса и бустинговые алгоритмы (GBM, XGBoost). В данной работе рассмотрен алгоритм CART (classification and regression trees). Отличительной особенностью этого алгоритма является то, что он предусматривает только два возможных варианта развития события, что подходит для реализации цели данного исследования. Основная идея CART заключается в разбиении первичного множества на два подмножества так, чтобы фирмы-банкроты находилась в одном множестве, а платежеспособные организации – в другом. Сложностью при использовании данного метода является определение момента остановки «расщепления множеств», так как возникает проблема переобучения. Выделяют следующие правила остановки:

- ✦ мера загрязненности меньше некоторого значения;
- ✦ ограничение на количество узлов или слоев дерева;
- ✦ размер родительского узла;
- ✦ размер узла-потомка.

Задание самих правил происходит с помощью кросс валидации. Несмотря на то, что существует ряд примеров успешного использования этого метода при прогнозировании дефолтов [5], данный метод имеет ряд недостатков: высокая чувствительность к входным данным, подверженность переобучению, сложность определения оптимальной архитекторы дерева.

Случайные леса. Случайные леса появились как модификация деревьев решений и, соответственно, часто позволяют получить более точные прогнозные результаты. Случайные леса состоят из заданного пользователем числа деревьев классификации, которые генерируются с использованием модифицированного алгоритма CART. Схема данного алгоритма представлена на *рисунке 3*. В алгоритме использованы два похода: каждое дерево обучается на своей подвыборке исходных данных (bootstrapped data); используются разные подмножества факторов при построении деревьев классификации. Данные действия приводят к построению, а затем и к «голосованию деревьев» относительно принадлежности объекта к определенному классу.

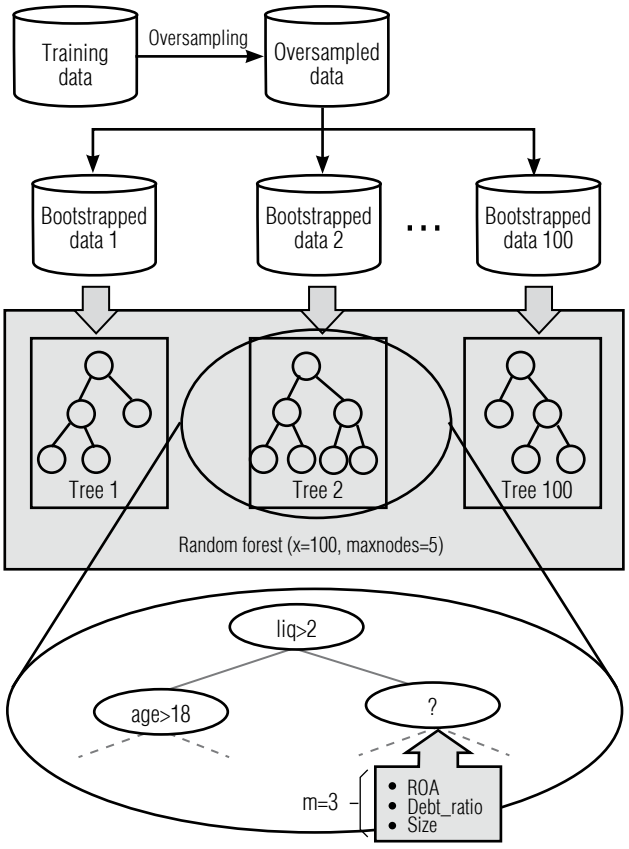


Рис. 3. Алгоритм случайного леса [4]

В отличии от регрессионных моделей, которые довольно чувствительны к выбросам, случайный лес (*RF*) является более робастным к данной проблеме. Преимуществом случайного леса является более высокая эффективность в случае несбалансированности данных (что актуально для поставленной нами задачи), а также меньшая подверженность переобучению. Недостатком алгоритма является меньшая прозрачность (в отличии от деревьев классификации) и, соответственно, более низкая интерпретация. Имеется относительная сложность в процессе определения параметров случайного леса. Определение параметров (количества деревьев, количества используемых при построении одного дерева факторов, максимального количества узлов в одном дереве) осуществлялось с помощью кросс-валидации.

Случайные леса часто используются для определения степени значимости какой-либо переменной. Идея оценки важности фактора основана на том, что перестановка значений важной переменной должна приводить к значительному увеличению частоты ошибок на тестовом множестве.

Искусственные нейронные сети. В настоящее время нейросетевое моделирование набирает популярность, особенно при прогнозировании явлений с однородными признаками. С помощью набора входных параметров выбирается архитектура сети, которая в наиболее простом варианте представлена тремя слоями. Первый слой содержит узлы (нейроны) для входных переменных (каждый нейрон имеет только один вход из внешней среды), второй слой содержит произвольное количество «скрытых» нейронов и поэтому называется скрытым слоем, третий слой содержит нейроны, отвечающие за результат. При этом в задачах прогнозирования банкротств последний слой содержит только один нейрон. Между входными и скрытыми нейронами задается связь с определенными весами. Например, для j-го нейрона в промежуточном слое и входными данными будет определена следующая линейная зависимость:

$$a_j = \sum_{i=1}^{N} \omega_{ij} \cdot x_i , \qquad (7)$$

где a_j – величина j-го нейрона;

ω_{ij} – вес j-го нейрона при переменной x_i.

Каждое значение a_j преобразуется с использованием некоторой активационной функции для получения фактического результирующего значения z_j нейрона j. Поскольку в нашем исследовании осуществляется прогнозирование двух классов, в качестве активационной функции удобно использовать логистическую функцию:

$$z_j = f(a_j) = \frac{1}{1+e^{-a_j}}, \qquad (8)$$

где z_j – нормированное значение величины j-го нейрона;

a_j – величина j-го нейрона.

Подобная процедура проводится для последующих слоев. Значения z_j снова взвешиваются, после чего преобразуются с использованием активационной функции для получения результата в конечном слое. Выделяют множество методов минимизации, идея которых состоит в том, что, начиная с исходного значения веса ω^0, генерируется последовательность векторов весовых коэффициентов ω^1, ω^2, ..., ω^k таких, что с каждой итерации алгоритма значение функции критерия качества уменьшается:

$$E\left(\omega^{k+1}\right) < E\left(\omega^k\right), \qquad (9)$$

где ω^k – значение веса после k-го шага обучения;

ω^{k+1} – значение веса после k+1-го шага обучения.

На *рисунке 4* представлен пример одной из итоговых моделей нейронной сети.

Одним из наиболее распространенных методов, используемых для обучения нейросетевых моделей, является метод наискорейшего спуска. В данном алгоритме корректировка весов выполняется в направлении максимального уменьшения критерия качества, т.е. в направлении, противоположному вектору градиента. Несмотря на то, что метод наискорейшего спуска сходится к оптимальному значению ω^* достаточно медленно, он является распространенным методом нахождения минимума во многих статистических библиотеках.

Одной из сложностей обучения нейронной сети является то, что функция критерия качества может иметь много локальных минимумов. В результате после инициализации модели можно прийти к локальному минимуму, что негативно скажется на результатах, получаемых на тестовом множестве. Для преодоления данной проблемы веса беспорядочно сортируются, а сам алгоритм обучения повторяется несколько раз. Оптимальные параметры (количество нейронов во внутреннем слое, количество внутренних слоев) так же, как и для деревьев классификации и случайных лесов были определены с помощью кросс-валидации.

Чтобы повысить эффективность нейронной сети, а также ускорить процесс обучения необходима предварительная обработка данных. Простой и эффективный шаг предварительной обработки включает масштабирование и центрирование данных.

4. Сравнительный анализ используемых моделей

В соответствии с используемой классификацией рассмотренные модели показали хорошее качество. В *таблице 2* приведена сортировка лучших моделей в группе в порядке убывания качества модели. Наилучшее качество показала искусственная нейронная сеть с одним скрытым слоем и четырьмя нейронами, с использованием алгоритма оверсемплинг. Отражено, что использование логистической модели с проведением дискретизации и переходом к WOE приводит к значительному росту точности моделей (коэффициент Джини в среднем увеличивается на 15%). Примечательно, что качество моделей соответствует уровню AUC в подобных работах [4, 5, 11].

Результаты однофакторного анализа с использованием регрессионных моделей свидетельствуют, что все рассмотренные факторы могут быть использованы при построении моделей бинарного

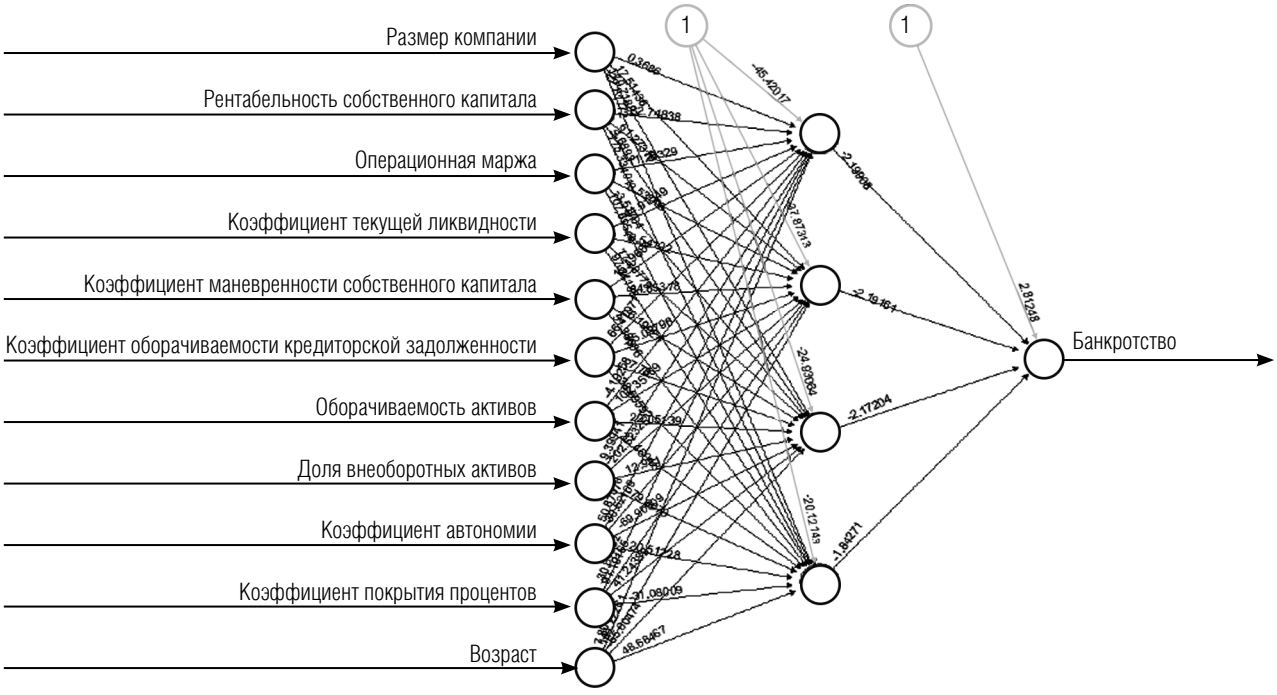


Рис. 4. Пример архитектуры нейронной сети

Таблица 2.

Оценка качества моделей на тестовом множестве

№	Модель	Коэффициент Джини, %	Доля неплатежеспособных компаний на тестовом множестве, %
1	Искусственные нейронные сети (oversample)	59,6	50
2	Искусственные нейронные сети	58,9	9,8
4	Логит–модель (oversample)	57,9	25
3	Пробит–модель (oversample)	57,6	20
5	Логит–модель	57,6	20
6	Логит–модель (undersample)	57,3	20
7	Искусственные нейронные сети (undersample)	56,0	50
8	Случайные леса (undersample)	52,4	15
9	Случайные леса	50,6	9,8
10	Случайные леса (oversample)	48,7	10
11	Деревья классификаций (oversample)	45,0	15
12	Логит–модель без дискретизации	42,2	9,8
13	Деревья классификации (со штрафами за неправильную классификацию миноритарного класса)	40,0	9,8
14	Деревья классификации	38,0	9,8
15	Деревья классификации (undersample)	38,0	50

выбора, так как для каждого из них значение AUC при однофакторном анализе выше 0,5. В итоговую многофакторную модель вошли восемь коэффициентов из шестнадцати факторов, отражающих разные аспекты рисков строительных компаний: ликвидность (коэффициент текущей ликвидности, коэффициент маневренности собственного капитала), рентабельность (рентабельность собственного капитала), платежеспособность (коэффициент покрытия процентов), оборачиваемость (оборачиваемость активов), деловая активность (доля внеоборотных активов), нефинансовые предикторы (возраст и размер компании). Включение в модель данных факторов приводит к повышению эффективности традиционных моделей (коэффициент Джини увеличивается с 0,38 до 0,58). При этом данные модели показали устойчивость к переобучению. В процессе многофакторного анализа гипотезы относительно знака зависимости вероятности неплатежеспособности от предполагаемых регрессоров были подтверждены. Значительных отличий в показателях точности между логит- и пробит-моделями обнаружено не было. Данный вывод согласуется со многими работами, поскольку функция логистического распределения и функция распределения стандартной нормальной случайной величины ведут себя примерно одинаково, а отличия связаны с более «тяжелыми хвостами» логистической функции распределения.

Ввиду устойчивости непараметрических алгоритмов к мультиколлиниарности, все рассмотренные ранее факторы использовались при построении моделей на основе методов машинного обучения. Анализ деревьев классификации и случайных лесов показал, что среди наиболее влиятельных факторов оказались коэффициент маневренности собственного капитала и коэффициент автономии (наибольшее падение индекса Джини в алгоритме случайного леса, первое разбиение в деревьях классификации). Это означает, что если у компании имеется значительный объем долговой нагрузки и она показывает негативный финансовый результат (в балансе собственный капитал принимает отрицательное значение), то это важный индикатор несостоятельности компании в следующий отчетный период. При этом нефинансовые факторы (возраст, размер компании) оказались практически несущественными, что отражено на *рисунке 5*. Таким образом, большой размер и длительный срок работы компании на рынке не могут гарантировать устойчивость на российском рынке.

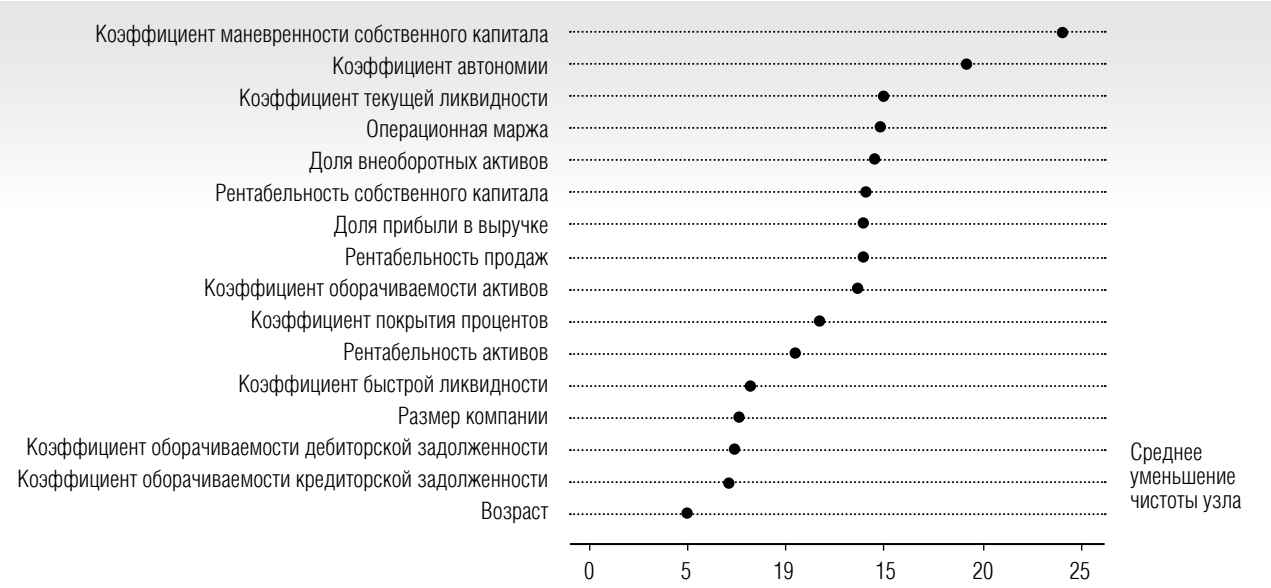


Рис. 5. Определение наиболее значимых параметров. Алгоритм случайных лесов

Динамика среднего значения коэффициента Джини в зависимости от доли неплатежеспособных компаний при оптимальных параметрах на обучающем множестве с использованием андерсемплинга и оверсемплинга (*рисунок 6*) показывает, что в задаче прогнозирования при данной структуре данных влияние доли неплатежеспособных компаний на обучающем множестве не оказывает значительного влияния на прогнозный потенциал того или иного метода. Данный вывод согласуется с работой Б.Б. Демешева и А.С. Тихоновой [14]. Отрицательная динамика метрики качества при росте доли неплатежеспособных компаний свидетельствует о смещении оценок при построении алгоритма (например, в алгоритме «случайный лес с использованием оверсемплинга»).

Применение метода борьбы с несбалансированностью зависит от типа используемой модели. Для логистической регрессии, искусственных нейронных сетей и деревьев классификации оверсемплинг показал более высокое качество. Однако использование оверсемплинга в методе случайных лесов приводит к переобучению. Поэтому для случайных лесов более высокую эффективность имеет андерсемплинг.

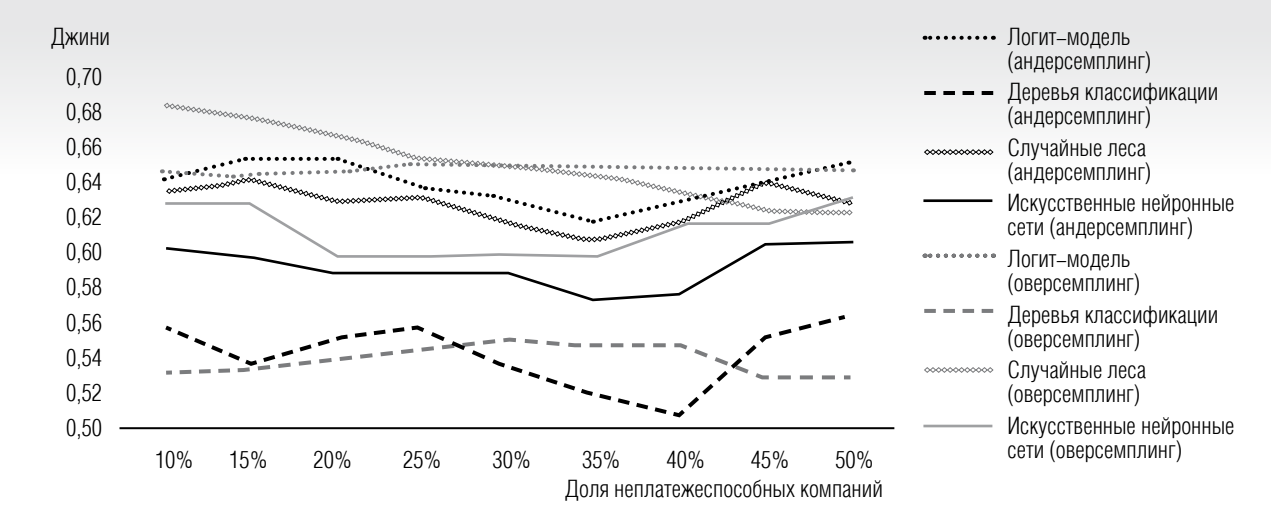


Рис. 6. Среднее значение коэффициента Джини на обучающем множестве

Заключение

Применение той или иной модели зависит от цели аналитика. В задачах прогноза нелинейные алгоритмы, как правило, показывают более высокий результат. Поэтому использование нейронных сетей и случайных лесов более приемлемо для данного типа задач. Однако эти модели проигрывают моделям бинарного выбора в затратах (временных, вычислительных) на расчеты, а также в интерпретации.

Рассмотренные нами алгоритмы показали приемлемое качество для использования в задачах прогнозирования банкротств строительных компаний. Как и ожидалось, наилучшей моделью оказалась искусственная нейронная сеть. Традиционные модели с дискретизацией показали хороший результат, при этом их результаты могут быть легко интерпретируемыми, а время расчетов минимально. Несмотря на достоинства деревьев классификации (легкость интерпретации, отсутствие ограничений на тип переменных, отсутствие необходимости задавать форму взаимосвязи в явном виде), этот алгоритм показал нестабильность и невысокую точность предсказаний.

В дальнейшем представляется перспективным включение в сравнение других нелинейных алгоритмов, например, моделей, основанных на бустинге (GBM, XGBoost), моделей опорных векторов и т.д. Более того, в данной работе в категорию банкротов попали компании, в отношении которых начата процедура легального банкротства, а также компании, ликвидировавшиеся добровольно. В дальнейшем представляется возможным разграничить данные категории с помощью единого федерального реестра сведений о банкротствах и выделить компании, в отношении которых началась процедура легального банкротства. Также представляется возможным провести межотраслевое сравнение рассмотренных методов, определить максимальный горизонт прогнозирования, при котором появляются признаки банкротств, провести диверсификацию внутри отдельных отраслей, а также использовать макроэкономические переменные при моделировании. ■

Литература

1. Карминский А.М., Костров А.В., Мурзенков Т.Н. Моделирование вероятности дефолта российских банков с использованием эконометрических методов / Препринт WP7/2012/04. Серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике». М.: ВШЭ, 2014.
2. Костров А.В. Сравнение статистических методов классификации для предсказания банкротства российских банков // Управление финансовыми рисками. 2016. Т. 47. № 3. С. 162–180.
3. Prediction of default probability for construction firms using the logit model / P. Tserng [et al.] // Journal of Civil Engineering and Management. 2014. Vol. 20. No 2. P. 247–255. DOI: 10.3846/13923730.2013.801886.
4. Behr A., Weinblat J. Default patterns in seven EU countries: A random forest approach // International Journal of the Economics of Business. 2017. Vol. 24. No 2. P. 181–222. DOI: 10.1080/13571516.2016.1252532.
5. Gepp A., Kumar K. Predicting financial distress: A comparison of survival analysis and decision tree techniques // Procedia Computer Science. 2015. Vol. 54. P. 396–404.
6. Tam K.Y., Kiang M.Y. Managerial applications of neural networks: the case of bank failure predictions // Management Science. 1992. Vol. 38. No 7. P. 926–947. DOI: 10.1287/mnsc.38.7.926.
7. Богданова Т.К., Шевгунов Т.Я., Уварова О.М. Применение нейронных сетей для прогнозирования платежеспособности российских предприятий обрабатывающих отраслей // Бизнес-информатика. 2013. № 2. С. 40–48.
8. Breiman L., Friedman J., Olshen R., Stone C. Classification and regression trees. Wadsworth, New York: Chapman and Hall, 1984.
9. Breiman L. Random forests // Machine Learning. 2001. Vol. 45. No 1. P. 5–32.
10. Odom M.D., Sharda R. A neural network model for bankruptcy prediction // International Joint Conference on Neural Networks. San Diego, USA, 17–21 June 1990. Vol. 2. P. 163–168. DOI: 10.1109/IJCNN.1990.137710.
11. Kumar P.R., Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review // European Journal of Operational Research. 2007. Vol. 180. No 1. P. 1–28. DOI: 10.1016/j.ejor.2006.08.043.
12. Trujillo-Ponce A., Samaniego-Medina R., Cardone-Riportella C. Examining what best explains corporate credit risk: accounting-based versus market-based models // Journal of Business Economics and Management. 2014. Vol. 15. No 2. P. 253–276. DOI: 10.3846/16111699.2012.720598.
13. Федеральный закон от 26.10.2002 № 127-ФЗ (ред. от 29.12.2017) «О несостоятельности (банкротстве)» (с изм. и доп., вступ. в силу с 28.01.2018).
14. Демешев Б.Б., Тихонова А.С. Прогнозирование банкротства российских компаний: межотраслевое сравнение / Препринт WP2/2014/04. Серия WP2 «Количественный анализ в экономике». М.: ВШЭ, 2014.
15. Siddiqi N. Credit risk scorecards: developing and implementing intelligent credit scoring. John Wiley & Sons, 2012.

Об авторах

Карминский Александр Маркович
доктор экономических наук, доктор технических наук;
ординарный профессор департамента финансов, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;
E-mail: karminsky@mail.ru
ORCID: 0000-0001-8943-4611

Бурехин Роман Николаевич
аспирант, аспирантская школа по экономике, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;
E-mail: romanvia93@yandex.ru
ORCID: 0000-0003-1130-0175

Comparative analysis of methods for forecasting bankruptcies of Russian construction companies

Alexander M. Karminsky
E-mail: karminsky@mail.ru

Roman N. Burekhin
E-mail: romanvia93@yandex.ru

National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

This paper is devoted to comparison of the capabilities of various methods to predict the bankruptcy of construction industry companies on a one-year horizon. The authors considered the following algorithms: logit and probit models, classification trees, random forests, artificial neural networks. Special attention was paid to the peculiarities of the training machine learning models, the impact of data imbalance on the predictive ability of models, analysis of ways to deal with these imbalances and analysis of the influence of non-financial factors on the predictive ability of models. In their study, the authors used non-financial and financial indicators calculated on the basis of public financial statements of the construction companies for the period from 2011 to 2017. The authors concluded that the models considered show acceptable quality for use in forecasting bankruptcy problems. The Gini or AUC coefficient (area under the ROC curve) was used as the quality markers of the model. It was revealed that neural networks outperform other methods in predictive power, while logistic regression models in combination with discretization follow them closely. It was found that the effective way to deal with the imbalance data depends on the type of model used. However, no significant impact on the imbalance in the training set predictive ability of the model was identified. The significant impact of non-financial indicators on the likelihood of bankruptcy was not confirmed.

Key words: bankruptcy; the construction sector; imbalance data; machine learning models; parametric models of prediction of bankruptcy.

Citation: Karminsky A.M., Burekhin R.N. Comparative analysis of methods for forecasting bankruptcies of Russian construction companies. *Business Informatics*, vol. 13, no 3, pp. 52–66. DOI: 10.17323/1998-0663.2019.3.52.66

References

1. Karminsky A.M., Kostrov A.V., Murzenkov T.N. (2012) *Approaches to evaluating the default probabilities of Russian banks with econometric methods*. Working paper WP7/2012/04 (Series "Mathematical methods of decision analysis in economics, business and policies"). Moscow: HSE (in Russian).
2. Kostrov A.V. (2016) Comparison of statistical classification methods to predict Russian banks failures. *Management of Financial Risks*, vol. 47, no 3, pp. 162–180 (in Russian).
3. Tserng P., Chen P.-C., Huang W.-H., Lei M.C., Tran Q.H. (2014) Prediction of default probability for construction firms using the logit model. *Journal of Civil Engineering and Management*, vol. 20, no 2, pp. 247–255. DOI: 10.3846/13923730.2013.801886.
4. Behr A., Weinblat J. (2017) Default patterns in seven EU countries: A random forest approach. *International Journal of the Economics of Business*, vol. 24. No 2. pp. 181–222. DOI: 10.1080/13571516.2016.1252532.
5. Gepp A., Kumar K. (2015) Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science*, vol. 54, pp. 396–404.
6. Tam K.Y., Kiang M.Y. (1992) Managerial applications of neural networks: the case of bank failure predictions. *Management Science*, vol. 38, no 7, pp. 926–947. DOI: 10.1287/mnsc.38.7.926.
7. Bogdanova T.K., Shevgunov T.Ya., Uvarova O.M. (2013) Using neural networks for predicting solvency of Russian companies on manufacturing industries. *Business Informatics*, no 2, pp. 40–48 (in Russian).
8. Breiman L., Friedman J., Olshen R., Stone C. (1984) *Classification and regression trees*. Wadsworth, New York: Chapman and Hall.
9. Breiman L. (2001) Random forests. *Machine Learning*, vol. 45, no 1, pp. 5–32.
10. Odom M.D., Sharda R. (1990) A neural network model for bankruptcy prediction. Proceedings of *International Joint Conference on Neural Networks. San Diego, USA, 17–21 June 1990*, vol. 2, pp. 163–168. DOI: 10.1109/IJCNN.1990.137710.
11. Kumar P.R., Ravi V. (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review. *European Journal of Operational Research*, vol. 180, no 1, pp. 1–28. DOI: 10.1016/j.ejor.2006.08.043.
12. Trujillo-Ponce A., Samaniego-Medina R., Cardone-Riportella C. (2014) Examining what best explains corporate credit risk: accounting-based versus market-based models. *Journal of Business Economics and Management*, vol. 15, no 2, pp. 253–276. DOI: 10.3846/16111699.2012.720598.
13. The Federal Law of 26 October 2002, No 127-FZ (as amended on 29 December 2017) "*On insolvency (bankruptcy)*" (in Russian).
14. Demeshev B.B., Tikhonova A.S. (2014) *Default prediction for Russian companies: intersectoral comparison*. Working paper WP2/2013/05 (Series WP2 "Quantitative Analysis of Russian Economy"). Moscow: HSE (in Russian).
15. Siddiqi N. (2012) *Credit risk scorecards: developing and implementing intelligent credit scoring*. John Wiley & Sons.

About the authors

Alexander M. Karminsky

Dr. Sci. (Econ.), Dr. Sci. (Tech.);

Professor, Department of Finance, National Research University Higher School of Economics,
20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: karminsky@mail.ru

ORCID: 0000-0001-8943-4611

Roman N. Burekhin

Doctoral Student, Doctoral School on Economics, National Research University Higher School of Economics,
20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: romanvia93@yandex.ru

ORCID: 0000-0003-1130-0175