

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Б.Б. Демешев, А.С. Тихонова

**ПРОГНОЗИРОВАНИЕ БАНКРОТСТВА
РОССИЙСКИХ КОМПАНИЙ:
МЕЖОТРАСЛЕВОЕ СРАВНЕНИЕ**

Препринт WP2/2014/04

Серия WP2

Количественный анализ в экономике

Москва 2014

Редактор серии WP2
«Количественный анализ в экономике»
В.А. Бессонов

Демешев, Б. Б., Тихонова, А. С.

Прогнозирование банкротства российских компаний: межотраслевое сравнение [Электронный ресурс]: препринт WP2/2014/04 / Б. Б. Демешев, А. С. Тихонова; Нац. исслед. ун-т «Высшая школа экономики». – Электрон. текст. дан. (2 Мб). – М.: Изд. дом Высшей школы экономики, 2014. – 27 с. – (Серия WP2 «Количественный анализ в экономике»).

Цель данной работы – сравнение подходов к моделированию критического финансового положения средних и малых российских непубличных компаний разных отраслей с помощью финансовых и нефинансовых показателей в 2011–2012 гг.

Используемые методы прогнозирования: логит- и пробит-модели, линейный дискриминантный анализ, квадратичный дискриминантный анализ, дискриминантный анализ смеси распределений, классификационное дерево и алгоритм случайного леса.

В исходной выборке содержится около миллиона наблюдений из базы данных RUSLANA, которые относятся к периоду 2011–2012 гг.

Вместо понятия банкротства мы используем понятие критического финансового положения, которое вынуждает компанию либо закрыться добровольно, либо быть ликвидированной по процедуре легального банкротства.

Исследуемые компании относятся к четырем отраслям: обрабатывающим производствам, операциям с недвижимостью, оптовой и розничной торговле и строительству. Сравнивая отрасли, мы приходим к нескольким важным выводам. С одной стороны, нужно строить отдельные модели для разных отраслей, поскольку разница между отраслями не может быть описана с помощью одного-двух дополнительных регрессоров в виде дамми-переменных.

С другой стороны, между отраслями много общего. Во-первых, сильно похожим между отраслями выходит ранжирование переменных по важности. Это, в частности, приводит к тому, что для всех четырех отраслей в качестве оптимального мы выбираем один и тот же набор регрессоров из рассматриваемых нами шести альтернатив. Вне зависимости от отрасли включение нефинансовых показателей улучшает прогнозную силу модели. Важными нефинансовыми переменными являются возраст компании и федеральный округ. Размер компании оказывает меньшее влияние, а организационная форма является самым слабым предиктором. Во-вторых, наилучшим алгоритмом стабильно оказывается случайный лес. Для всех отраслей у случайного леса показатель качества прогнозов (площадь под ROC-кривой) достигает примерно $\frac{3}{4}$.

Задача прогнозирования дефолта предприятия на следующий год интересна как банкам и другим кредиторам предприятий, так и государственным контролирующим органам.

Ключевые слова: прогнозирование банкротства; сравнение моделей; средние и малые предприятия; оптовая и розничная торговля; обрабатывающая промышленность; операции с недвижимостью; строительство.

JEL: C14, C45, G30, G33

Демешев Борис Борисович – старший преподаватель, кафедра математической экономики и эконометрики, департамент прикладной экономики, НИУ ВШЭ; 119049, Москва, Российская Федерация, ул. Шаболовка, 28, каб. 112; E-mail: bdemeshev@hse.ru; тел.: 8 903 287 34 22

Тихонова Анна Сергеевна – студентка 1-го года магистерской программы «Финансовая экономика» МИЭФ, НИУ ВШЭ; 119049, Москва, Российская Федерация, ул. Шаболовка, 28, каб. 112; E-mail: annette.tikhonova@gmail.com; тел.: 8 910 417 17 07

Авторы статьи выражают благодарность Ершову Э.Б. и Канторовичу Г.Г. за сделанные в ходе научно-исследовательского семинара комментарии и замечания.

Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу: <http://www.hse.ru/org/hse/wp>

© Демешев Б. Б., 2014

© Тихонова А. С., 2014

© Оформление. Издательский дом
Высшей школы экономики, 2014

1. Введение

В последнее время тема банкротств предприятий привлекает внимание исследователей и становится всё более актуальной в связи с недавним кризисом 2008 – 2009 гг. и с новыми требованиями, предъявляемыми Базельскими соглашениями. До 2007 г. в Базельских соглашениях не было значительных отличий в требованиях к капиталу компаний крупного бизнеса по сравнению с компаниями среднего и малого бизнеса. Тем не менее внесённые поправки вынудили банки строить отдельные модели для компаний, относящихся к различным категориям по размеру.

Цель данной работы – сравнить подходы к моделированию критического финансового положения средних и малых российских непубличных компаний разных отраслей с помощью финансовых и нефинансовых показателей в 2011 – 2012 гг. Мы исследуем четыре отрасли: обрабатывающие производства, операции с недвижимостью, оптовую и розничную торговлю, строительство. Для достижения цели необходимо ответить на ряд вопросов:

- какой метод прогнозирования лучше остальных?
- улучшает ли добавление нефинансовых характеристик компаний прогнозную силу моделей?
- различаются ли модели по отраслям и организационным формам компаний?

Построение скоринговых моделей для крупных фирм основано на иных принципах, чем построение моделей банкротства для более мелких фирм. Так, например, крупные компании котируются на бирже, и у них другие требования по раскрытию информации. Однако именно компании среднего и малого бизнеса способствуют инновациям и создают прочную основу экономики каждой страны. Поэтому банки и иные кредитные организации, принимающие решение о предоставлении кредита фирмам, и государственные контролирующие органы нуждаются в моделях с высокой прогнозной силой.

Новизна работы состоит в следующем. Впервые в одной работе сравнивается большое количество статистических методов: логит- и пробит-модели, линейный дискриминантный анализ (ЛДА), квадратичный дискриминантный анализ (КДА), дискриминантный анализ смеси распределений (СДА), метод классификационных деревьев и алгоритм случайного леса. Впервые оценивается столь большой массив данных российских предприятий. В исходной выборке содержится 950 тыс. наблюдений. Впервые на российских данных сделана попытка оценить и учесть неоднородность по отраслям и формам организации предприятия. Период анализа – 2011 – 2012 гг. Мы расширяем понятие банкротства до понятия критического финансового положения, которое несовместимо с дальнейшим функционированием компании. Исследуются три вида компаний: (1) действующие в данный момент компании; (2) неактивные, а именно ликвидированные в результате банкротства, и (3) добровольно ликвидированные компании. Важно отметить, что используется отчётность компаний, адаптированная к международным стандартам финансовой отчётности.

2. Обзор литературы

В современном мире различия в функционировании компаний разного размера огромны. Небольшим компаниям сложнее взять кредит из-за ограничения на допустимый размер долга. В работе [Kolari, Ou, Shin, 2006] утверждается, что некрупные компании рискуют гораздо больше, чем крупные. Некоторые исследователи считают, что нельзя проводить сравнительный анализ крупных и небольших компаний [Sirirattanaphonkun, Pattarathammas, 2012]. Стандартное деление фирм на группы предусматривает две категории: (1) крупные; (2) средние и малые. Компании второй группы можно разделить далее на три подгруппы: микро-, малые и средние предприятия. Однако универсальных критериев для отнесения компании к той или иной группе не существует.

В нашем исследовании для разделения компаний на различные группы мы используем параметры численности персонала (number of employees) и выручки (turnover). Федеральный закон № 209-ФЗ «О развитии малого и среднего предпринимательства в Российской Федерации» утверждает границы для этих параметров. Таблица 1 содержит границы каждого показателя. Также для сравнения мы приводим значения этих же параметров для европейских предприятий [European Commission. Enterprise and Industry. Small and medium-sized enterprises: SME Definition; Basel II]. Фирма, относящаяся к определённому размеру, должна соответствовать сразу обоим критериям.

Таблица 1. Критерии определения размера компании

Размер	Численность персонала (чел.)		Выручка	
	Европейский союз	Россия	Европейский союз	Россия
Микро-	1–10	1–15	≤ €2 млн	≤ €1,4 млн (RUB 60 млн)
Малое	11–50	16–100	≤ €10 млн	≤ €9,6 млн (RUB 400 млн)
Среднее	51–250	101–250	≤ €50 млн	≤ €24 млн (RUB 1 млрд)

В числе требований, которые накладывают государство и институциональная среда на компании малого и среднего бизнеса, можно выделить условие повышенного обеспечения долга, а также повышенные ставки по кредитам [Financing SMEs and Entrepreneurs 2013: An OECD Scoreboard Final Report, 2013]. По этим причинам компании данного размера более подвержены финансовой нестабильности, и раннее обнаружение признаков, указывающих на возможные финансовые трудности, крайне необходимо для принятия своевременных мер.

Определять дефолт можно по-разному, наиболее широко распространены лишь два определения. Первый вариант подразумевает под дефолтом неспособность оплатить проценты по долгу или часть основного тела долга [Maleev, Nikolenko, 2010]. Второй вариант (чаще встречающийся) подразумевает под дефолтом легальное банкротство [Hunter, Isachenkova, 2001; Ciampi, Vallini, Gordini, 2009; Khorasgani, 2011]. Он учитывает, что не всякая невозможность платить может считаться банкротством.

Тема прогнозирования дефолта компаний начала привлекать внимание исследователей ещё в прошлом веке: в работе [Beaver, 1966] предложено использовать одномерный параметрический метод, главным недостатком которого является выбор порога отсечения. В работе [Altman, 1968] для задач такого рода был впервые использован линейный дискриминантный анализ (ЛДА). Наличие наперёд заданных финансовых отношений не позволяет принять во внимание все источники дохода определённой фирмы. К тому же наличие нефинансовых характеристик компаний в модели во многом определяет качество прогнозов модели [Lugovskaya, 2010].

Позже в работе [Wei, Li, Chen, 2007] появилась критика использования ЛДА в данной ситуации: алгоритм ЛДА может ошибочно классифицировать исходы, потому что ковариационные матрицы банкротов и активных компаний, вероятно, не совпадают. Применимость ЛДА и КДА (квадратичного дискриминантного анализа) оказывается также под вопросом в силу того, что финансовые отношения часто не имеют нормального распределения [Ohlson, 1980; Wilson, Sharda, 1994].

Важным этапом в развитии прогнозирования вероятности дефолта стало использование логит- и пробит-моделей [Martin, 1977; Ohlson, 1980]. Оказалось, что дискриминантный анализ (ЛДА и КДА) уступает логит- и пробит-моделям по прогнозной силе.

Среди других методов прогнозирования дефолта можно назвать метод опорных векторов [Härdle et al., 2007] и нейронные сети [Tam, Kiang, 1992; Altman, Marco, Varetto, 1994].

Однако все методы имеют недостатки. Логит-, пробит-модели и ЛДА требуют добавления переменных, чтобы получить немонотонную зависимость между дефолтом и объясняющими переменными. Метод опорных векторов, нейронные сети и классификационные деревья подвержены проблеме сверхподгонки.

Деятельность компании характеризуется множеством финансовых показателей. Выбор финансовых отношений для прогнозирования банкротства не является простой задачей. Чаще всего коэффициенты делят на пять групп: рентабельность, ликвидность, оборачиваемость, финансовый рычаг, обслуживание долга. Так, в работе [Altman, Sabato, 2007], посвящённой исследованию предприятий малого и среднего бизнеса, в качестве основных рассматриваются следующие показатели: отношение прибыли с учётом процентных платежей, налогов и амортизации к суммарным активам ($EBITDA / Total\ assets$), отношение краткосрочного долга к собственному капиталу ($Short-term\ debt / Total\ equity$), отношение нераспределённой прибыли к суммарным активам ($Retained\ earnings / Total\ assets$), отношение наличности к суммарным активам ($Cash / Total\ assets$), отношение прибыли с учётом процентных платежей, налогов и амортизации к процентным платежам ($EBITDA / Interest\ expenses$).

Помимо размера компании существенными нефинансовыми показателями являются также возраст компании, организационная форма и отрасль, в которой компания работает. Некоторые исследователи, в частности [Pompe, Bilderbeek, 2005], приходят к выводу, что прогнозировать вероятность банкротства давно существующих фирм проще, чем молодых компаний. Они предлагают оценивать различные модели по возрастным категориям. В работе [Falkenstein, Boral, Carty, 2000] утверждается, что связь между финансовыми отношениями и вероятностью банкротства различна для публичных и непубличных компаний. В исследованиях [Zeitun, Tian, Keen, 2007; Kaplinski, 2008] отмечается, что методы необходимо адаптировать в зависимости от отрасли.

Анализу компаний среднего и малого бизнеса посвящены, например, работы [Edmister, 1972, Altman, Sabato, 2007; Ciampi, Vallini, Gordini, 2009].

Дефолту российских компаний уделено существенно меньше исследовательского внимания, чем дефолту иностранных компаний. В 1990-е годы в связи со значительными изменениями в законодательстве и правилах ведения бухгалтерского учёта исследования подобного рода были практически невозможны. По данным середины 1990-х годов была написана работа, в которой сравнивались банкротства российских и британских фирм [Hunter, Isachenkova, 2001]. Попытки Зайцевой, Сайфуллина и Кадикова адаптировать модели Альтмана и Олсона не увенчались успехом.

Алгоритм ЛДА для прогнозирования вероятности банкротства на российских данных используется в работах [Lugovskaya, 2010; Жданов, Афанасьева, 2011]. Во второй работе [Жданов, Афанасьева, 2011] в дополнение к ЛДА оцениваются логит-модели, которые наравне с пробит-моделями используются в работах [Фёдорова, Гиленко, Довженко, 2012; Makeeva, Neretina, 2013]. В работе [Макеева, Бакурова, 2006] применяются нейронные сети.

В основном исследователи концентрируются на одной отрасли. Среди исключений можно отметить работу [Lugovskaya, 2010], где в том числе учтён размер компаний и рассмотрены несколько отраслей. Количество наблюдений во многих работах очень небольшое. Кроме того, в большинстве работ правовые формы никак не обозначены. Выбор финансовых переменных на основании зарубежных исследований зачастую слабо обоснован, потому что отчётность по РСБУ отличается от отчётности по МСФО.

3. Описание данных

Мы используем данные по российским компаниям, которые были собраны из базы данных российских, украинских и казахских компаний РУСЛАНА [ruslana.bvdep.com]. Период наблюдения: 2011 – 2012 гг. Данные годовые.

Мы анализируем только непубличные российские компании, которые относятся к среднему и малому бизнесу и к одной из четырёх отраслей: обрабатывающим производствам, операциям с недвижимостью, оптовой и розничной торговле, строительству (по ОКВЭД). Исследуемые фирмы имеют микро-, малый или средний размер, который соответствующий критериям Федерального закона РФ № 209-ФЗ «О развитии малого и среднего предпринимательства в Российской Федерации». Непубличные компании представлены обществами с ограниченной ответственностью (ООО) и закрытыми акционерными обществами (ЗАО).

Мы обобщаем категорию компаний легальных банкротов до категории компаний, находящихся в критическом финансовом положении. Под этим понятием мы подразумеваем и компании, в отношении которых была начата процедура легального банкротства, и компании, ликвидированные добровольно. Эти две категории имеют много общего, оба типа компаний не могут продолжать свою деятельность из-за критических финансовых трудностей. Разница между ними состоит в том, что одни накопили такой объём долга, что не могут его покрыть, а вторые столкнулись с такими большими убытками, что наиболее выгодным решением является закрытие бизнеса.

В табл. 2 представлено краткое описание процедур банкротства и добровольной ликвидации. Более подробно процедуры можно посмотреть в Федеральных законах

№ 127-ФЗ «О несостоятельности (о банкротстве)» и № 129-ФЗ «О государственной регистрации юридических лиц и индивидуальных предпринимателей».

Таблица 2. Процедуры банкротства и добровольной ликвидации в России

Процедура банкротства	Процедура добровольной ликвидации
Признаки банкротства	Причины
Финансовое оздоровление	Решение уполномоченного органа организации
Внешнее управление	Проверка и расследование
Признание банкротства	Рассмотрение судом
Конкурсное производство	Продажа имущества
Ликвидация как результат банкротства	Ликвидация как результат добровольного решения

Таким образом, в выборке содержатся три типа фирм в зависимости от статуса: активные, добровольно ликвидированные компании и ликвидированные банкроты.

Все переменные в данной работе делятся на два класса: финансовые отношения и нефинансовые характеристики. Финансовые отношения мы рассчитали на основании финансовой отчётности: баланса и отчёта о прибылях и убытках. Множество финансовых отношений мы делим на пять групп на основе изученной литературы и собственного анализа:

1. Показатели рентабельности. Данные отношения показывают, в состоянии ли компания покрывать свои издержки, а также получать прибыль. Показатели рентабельности могут быть и положительными, и отрицательными. Для наглядности приведём на рис. 1 гистограммы рентабельности активов *roa* (Net income / Total assets).

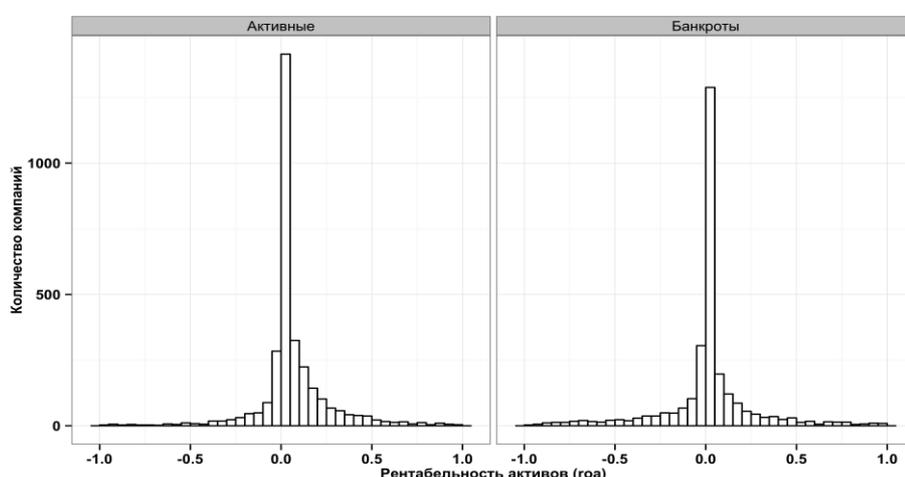


Рис. 1. Распределение рентабельности активов в зависимости от статуса

2. Показатели ликвидности. Данные отношения показывают, сколько имеется у компании ликвидных и неликвидных активов, а также время, необходимое для реализации этих активов. Для примера на рис. 2 приведём показатель ликвидности lr ((Current assets – Stocks) / Current liabilities).

3. Показатели финансового рычага. Данные отношения показывают возможный рост прибыли, связанный с увеличением долга. Более того, они характеризуют устойчивость компании, отражая соотношение собственных и заёмных средств.

4. Показатели обслуживания долга. Данные отношения характеризуют кредитную историю компании. Они позволяют сделать вывод, насколько добросовестно и оперативно компания совершает выплаты по своим долгам.

5. Показатели деловой активности. Данные отношения показывают скорость оборачиваемости средств компании, определяя необходимость в оборотном капитале.

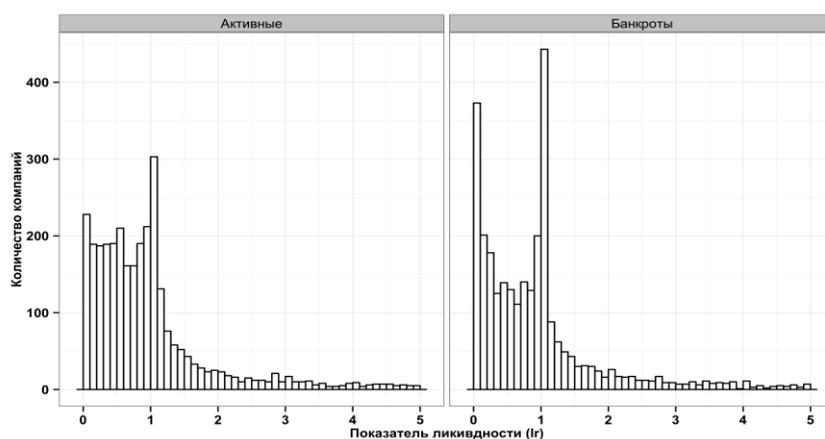


Рис. 2. Распределение показателя ликвидности lr в зависимости от статуса

Как видно на рис. 1 и 2, распределение финансовых отношений в наших данных не является нормальным и имеет очень тяжёлые хвосты. Мы приводим графики для выборки из всех отраслей в целом, так как распределение слабо отличается между отраслями.

Список нефинансовых показателей представлен в табл. 3.

Таблица 3. Описание переменных

Переменная	Описание
Организационная форма	Закрытое акционерное общество Общество с ограниченной ответственностью
Статус	Активное Добровольно ликвидированное Ликвидированное в результате банкротства
Дата статуса	Дата ликвидации (если предприятие было ликвидировано)
Возраст	Возраст компании в годах
Дата создания	Если дата не содержала дня, то ставится середина месяца (гггг-мм-15). Если дата не содержала ни дня, ни месяца, то ставится середина года (гггг-07-01)
Дефолт	Если компания обанкротится в текущем году – 1 Если компания не обанкротится в текущем году – 0
Дефолт в следующем году	Если компания обанкротится в следующем году – 1 Если компания не обанкротится в следующем году – 0
Федеральный округ	Укрупнение административного деления до федеральных округов
Последний доступный размер	Микро-, малое и среднее

Особые сложности возникли с определением размера компании. Во-первых, мы использовали два критерия – численность персонала и выручку компании, поэтому в некоторых случаях эти два критерия относили одну и ту же компанию к разным категориям размера. В таком случае мы считали количество работников приоритетным критерием. Во-вторых, в данных было много пропусков. Если пропуски были только по одному критерию, то мы использовали второй критерий. Если пропуски были по обоим критериям, то мы определяли размер по прошлому году, если же и в прошлом году не было данных по обоим критериям, то мы брали позапрошлый год. Распределение компаний по размеру для разных отраслей и статусов представлено на рис. 3.

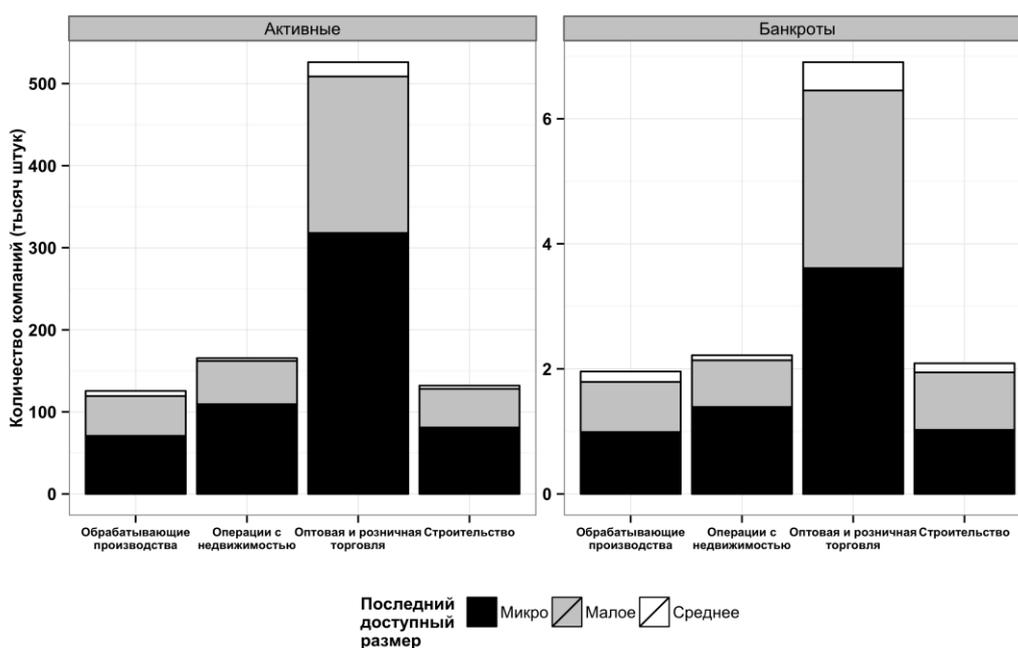


Рис. 3. Распределение компаний по размеру

На рис. 3 видно, что наибольшее число компаний относится к отрасли оптовой и розничной торговли, а остальные три отрасли примерно равны по численности. Распределение размера практически одинаково для активных компаний и банкротов. По абсолютной величине количество банкротов примерно в 100 раз меньше, чем активных предприятий. Отметим, что средних предприятий существенно меньше, чем микро- и малых.

4. Используемые алгоритмы

Выбранные нами алгоритмы оценивания можно разделить на две группы:

1. Вероятностные модели, оцениваемые с помощью метода максимального правдоподобия. Это полностью параметрические вероятностные модели, в которых точно специфицируется закон распределения зависимой переменной, индикатора дефолта в следующем году, при фиксированных значениях регрессоров. К этой группе относятся линейный (ЛДА) и квадратичный дискриминантный

анализ (КДА), дискриминантный анализ смеси распределений (СДА), логит- и пробит-модели.

2. Алгоритмы, основанные на бинарных классификационных деревьях. Здесь нет чёткого множества оцениваемых параметров. Данный тип моделей предполагает полностью нелинейную зависимость индикатора дефолта от регрессоров. К этой группе относятся алгоритм случайного леса и собственно бинарное классификационное дерево.

Все методы предполагают некоторую зависимость вероятности дефолта в следующем 2012 г. от показателей 2011 г.:

$$(1) \quad P(def_{i,2012} = 1) = f(x_{i,2011}),$$

где $x_{i,2011}$ – вектор характеристик компании i в 2011 г., а $def_{i,2012}$ – индикатор дефолта компании i в 2012 г. Далее индекс времени мы будем опускать.

Три метода дискриминантного анализа предполагают, что для случайно выбранного предприятия существует неизвестная априорная вероятность p быть банкротом. Далее методы дискриминантного анализа постулируют закон распределения регрессоров при каждом статусе предприятия.

В линейном дискриминантном анализе (ЛДА) предполагается, что при фиксированном статусе предприятия функция совместное распределение регрессоров является многомерным нормальным с постоянной ковариационной матрицей:

$$(2) \quad f(x_i | def_i = 0) \sim N(\mu_0, \Sigma)$$

$$(3) \quad f(x_i | def_i = 1) \sim N(\mu_1, \Sigma)$$

Квадратичный дискриминантный анализ (КДА) является обобщением линейного и предполагает, что распределения при разных значениях статуса def_i могут отличаться не только вектором математических ожиданий μ , но и ковариационной матрицей Σ :

$$(4) \quad f(x_i | def_i = 0) \sim N(\mu_0, \Sigma_0)$$

$$(5) \quad f(x_i | def_i = 1) \sim N(\mu_1, \Sigma_1)$$

Дискриминантный анализ смеси распределений (СДА) по-другому обобщает линейный дискриминантный анализ. Ковариационная матрица Σ предполагается одинаковой, а закон распределения регрессоров при фиксированном статусе предполагается смесью трёх многомерных нормальных распределений

$$(6) \quad f(x_i | def_i = 0) \sim p_{0a}N(\mu_{0a}, \Sigma) + p_{0b}N(\mu_{0b}, \Sigma) + p_{0c}N(\mu_{0c}, \Sigma)$$

$$(7) \quad f(x_i | def_i = 1) \sim p_{1a}N(\mu_{1a}, \Sigma) + p_{1b}N(\mu_{1b}, \Sigma) + p_{1c}N(\mu_{1c}, \Sigma)$$

В логит- и пробит-моделях предполагается существование скрытой склонности предприятия к банкротству, def_i^* , которая зависит от регрессоров линейным образом:

$$(8) \quad def_i^* = \alpha + x_i\beta + \varepsilon_i$$

Предприятие оказывается банкротом, если $def_i^* > 0$, и активным, если $def_i^* \leq 0$. Случайные ошибки ε_i предполагаются независимыми и одинаково распределёнными. В логит-модели они имеют логистическое распределение, а в пробит-модели – стандартное нормальное.

Алгоритм бинарного классификационного дерева состоит в пошаговом построении дерева. На каждом шаге очередной узел дерева делится на два подузла. Для деле-

ния узла на два используется та переменная и то её пороговое значение, которые обеспечивают максимальное падение *индекса Джини*, $\Delta I_G = I_{G1} - I_{G0}$, при делении узла.

Индексом Джини заданного узла дерева до деления, I_{G0} , называется вероятность несовпадения статусов двух случайно выбираемых из данного узла предприятий. То есть

$$(9) \quad I_{G0} = 2h(1 - h),$$

где h – доля предприятий-банкротов в заданном узле.

Индексом Джини заданного узла дерева после его деления на два подузла называется вероятность несовпадения статусов двух случайно выбираемых предприятий, если первое случайным образом выбирается из обоих подузлов, а второе – из того же подузла, что и первое. То есть

$$(10) \quad I_{G1} = \frac{n_L}{n} 2h_L(1 - h_L) + \frac{n_R}{n} 2h_R(1 - h_R),$$

где n – количество предприятий в узле до разбиения на два подузла, n_L и n_R – количества предприятий в левом и правом подузлах соответственно, а h_L и h_R – доля предприятий-банкротов в левом и правом подузлах соответственно.

Построение дерева заканчивается, когда достигается заданная сложность дерева. Существует много способов измерить сложность дерева, в нашем алгоритме в каждом терминальном узле должно быть не менее 5 предприятий, а высота дерева (длина самой длинной ветви) не может быть больше 31.

Алгоритм случайного леса состоит в построении 500 классификационных деревьев. Каждое дерево строится по своей случайной подвыборке. Подвыборка для каждого дерева выбирается случайным образом с возвращениями из исходной выборки и имеет такой же размер. Следует отметить две особенности построения деревьев в алгоритме случайного леса:

- каждое дерево строится до идеальной классификации, т.е. пока в терминальном узле не окажется ровно одно предприятие;
- во время построения одного дерева перед каждым делением узла на два из всего множества k имеющихся регрессоров предварительно случайным образом отбирается $\lfloor \sqrt{k} \rfloor$ регрессоров. А затем уже из этих $\lfloor \sqrt{k} \rfloor$ предварительно отобранных регрессоров выбирается тот регрессор и такой порог, которые обеспечивают максимальное падение индекса Джини.

Отметим, что эти семь моделей оцениваются отдельно для каждой из четырёх отраслей по двум видам выборок – *балансированной* и *небалансированной*.

Балансированной мы называем выборку, в которой равное количество активных предприятий и предприятий банкротов. Для балансировки выборки мы отбираем все дефолтные предприятия (так как их меньше, чем активных) и равное количество активных предприятий, выбираемых из данной отрасли случайным образом.

В *небалансированную* выборку мы также включаем все дефолтные предприятия и 20% всех активных предприятий из данной отрасли, отобранных случайным образом.

5. Выбор объясняющих переменных

Проблема выбора объясняющих переменных стоит перед каждым исследователем, однако в данной теме она особенно сложна в силу огромного числа всевозможных финансовых отношений, которые можно рассчитать с помощью отчётности компании. В данной работе мы применяем *три способа отбора финансовых переменных* и *два набора нефинансовых характеристик*.

Базовые версии моделей включают в себя финансовые переменные, отобранные с помощью одного из представленных ниже способов, и одну нефинансовую характеристику компании – её возраст.

Первый способ выбора финансовых отношений заключается в отборе тех показателей, которые наиболее близки к финансовым переменным в модели Альтмана и Сабато [Altman, Sabato, 2007] для компаний среднего и малого размера. Также к ним добавлена переменная *sr*, которая отражает степень стабильности компании. Базовая версия модели с этим набором показателей выглядит так:

$$(11) \quad P_i = f(iptd_i, ebta_i, stdte_i, roa_i, liq_i, sr_i, age_i)$$

Второй способ выбора финансовых отношений заключается в отборе переменных, которые наиболее часто использовались в исследованиях по прогнозированию банкротства компаний в мире. Мы взяли обзор такого рода исследований с 1930 по 2007 г. [Bellovary, Giacominio, Akers, 2007]. Если переменная упоминалась от 27 до 54 раз (максимум, который указан в статье), то мы её включали в модель. Базовая версия модели с этим набором показателей выглядит так:

$$(12) \quad P_i = f(iptd_i, roa_i, cr_i, wcta_i, ebta_i, lr_i, tdt_a_i, age_i)$$

Третий способ выбора финансовых отношений и нефинансовых переменных одновременно заключается в отборе переменных с помощью метода LASSO. Базовая версия модели с этим набором показателей выглядит так:

$$(13) \quad P_i = f(gg_i, nat_i, ebtm_i, ltdta_i, wcta_i, age_i),$$

где $P_i = P(def_i = 1)$.

Расширенные версии моделей подразумевают тот же набор финансовых переменных и расширенный набор нефинансовых переменных: в модель включается не только возраст, но и последний доступный размер компании, федеральный округ и организационная форма. Все три версии представлены ниже.

Расширенная модель Альтмана и Сабато:

$$(14) \quad P_i = f(iptd_i, ebta_i, stdte_i, roa_i, liq_i, sr_i, age_i, fedreg_i, lasize_i, legal_form_i)$$

Расширенная модель с популярными финансовыми показателями:

$$(15) \quad P_i = f(iptd_i, roa_i, cr_i, wcta_i, ebta_i, lr_i, tdt_a_i, age_i, fedreg_i, lasize_i, legal_form_i)$$

Расширенная модель с переменными, отобранными с помощью LASSO:

$$(16) \quad P_i = f(gg_i, nat_i, ebtm_i, ltdta_i, wcta_i, age_i, fedreg_i, lasize_i, legal_form_i),$$

где обозначение вида *fedreg_i* означает добавление в модель соответствующего числа дамми-переменных.

В табл. 4 представлены переменные обоих типов (финансовые и нефинансовые), их формулы и указание, в какую модель они входят.

Таблица 4. Переменные в моделях

Название в данных	Формула	Альтман		Популярные		LASSO	
		Возраст	Все	Возраст	Все	Возраст	Все
stdte	$\frac{\text{Short - term debt}}{\text{Total equity}}$	+	+				
liq	$\frac{\text{Cash}}{\text{Total assets}}$	+	+				
sr	$\frac{\text{Total assets}}{\text{Total equity}}$	+	+				
ebta	$\frac{\text{Total assets}}{\text{EBIT}}$	+	+	+	+		
roa	$\frac{\text{Total assets}}{\text{Net income}}$	+	+	+	+		
iptd	$\frac{\text{Total debt}}{\text{Interest paid}}$	+	+	+	+		
cr	$\frac{\text{Current liabilities}}{\text{Current assets - stocks}}$			+	+		
lr	$\frac{\text{Current liabilities}}{\text{Total debt}}$			+	+		
tdta	$\frac{\text{Total assets}}{\text{Working capital}}$			+	+		
wcta	$\frac{\text{Total assets}}{\text{Non - current liabilities + loans}}$			+	+	+	+
gg	$\frac{\text{Total equity}}{\text{Turnover}}$					+	+
nat	$\frac{\text{Total equity + Non - current liabilities}}{\text{EBIT}}$					+	+
ebtm	$\frac{\text{Turnover}}{\text{Long - term debt}}$					+	+
ltdta	$\frac{\text{Total assets}}$					+	+
age	Возраст компании	+	+	+	+	+	+
fedreg	Федеральный округ		+		+		+
lasize	Размер компании		+		+		+
legal_form	Организационная форма		+		+		+

Всего получилось шесть формул, которые оценивались с помощью семи методов по двум видам выборок: балансируемым и небалансируемым. Следует подчеркнуть важный момент. Три метода дискриминантного анализа предполагают, что регрессоры имеют многомерное нормальное распределение (или смесь многомерных нормальных), поэтому использование факторных переменных в качестве регрессоров явно нарушает предпосылки этих методов. Поэтому расширенные версии формул, включающие дамми-переменные, анализировались с помощью четырёх алгоритмов (классификационные деревья, алгоритм случайного леса, логит- и пробит-модели), а базовые версии формул анализировались с помощью всех семи алгоритмов. Таким образом, всего было оценено 66 моделей.

Количество наблюдений отличается в зависимости от вида выборки (балансируемая или небалансируемая) и выбранного набора объясняющих переменных (в силу наличия пропусков). Количество наблюдений для каждого случая представлено в табл. 5.

Таблица 5. Количество наблюдений в моделях

Отрасль	Тип выборки	Набор регрессоров		
		Альтман	Популярные	LASSO
Обрабатывающие производства	Балансированные	774	800	388
	Небалансированные	10 946	10 883	7 533
Операции с недвижимостью	Балансированные	988	1 002	484
	Небалансированные	11 977	11 906	7 615
Оптовая и розничная торговля	Балансированные	3 986	3 892	3 274
	Небалансированные	44 239	43 258	34 430
Строительство	Балансированные	770	806	362
	Небалансированные	10 533	10 478	6 939

6. Эмпирические результаты

6.1. Критерий отбора методов и наборов регрессоров

Для выбора «наилучшего» алгоритма оценивания и «наилучшего» набора регрессоров важно определить, что означает «лучше». Следует отметить, что задача прогнозирования принципиально отличается от определения влияния переменных. Например, может случиться так, что в модель, имеющую наименьшую среднеквадратическую ошибку прогноза, не будут включены значимые переменные. В нашем случае зависимая переменная является бинарной, банкрот или активное предприятие, поэтому критерий «лучше» должен прямо или косвенно основываться на количестве верно классифицированных предприятий.

Самый простой критерий качества, т.е. количество верно классифицированных предприятий, к сожалению, непригоден. Действительно, при таком критерии качества даже самая тривиальная модель «все предприятия в следующем году будут активными» получит очень высокую оценку, ведь доля верно угаданных статусов предприятий окажется примерно 0,99. В нашем случае мы стоим перед выбором между долей верно классифицированных активных предприятий (специфичность) и долей верно классифицированных банкротов (чувствительность). Увеличивая порог для прогнозной вероятности, за которым предприятие признаётся банкротом, мы снижаем чувствительность и увеличиваем специфичность. Поэтому мы выбрали интегральный показатель качества – площадь под ROC-кривой, т.е. площадь под графиком зависимости чувствительности от доли ошибочно признанных банкротами (единица минус специфичность).

В литературе, посвящённой банкротству, встречаются прогнозы разных типов. Некоторые авторы оценивают качество прогнозов по той же выборке, по которой оценивались модели. Мы считаем данный подход некорректным, так как он приводит к существенному завышению показателей прогнозной силы. Более корректным является оценивание качества прогнозов вне той выборки, по которой производилось оценивание. Здесь также возможны два варианта. Можно поделить выборку 2011 г. на две части, по одной части оценивать модель, а по второй оценивать качество прогнозов. Можно оценивать модель по данным 2011 г., а прогнозировать используя данные по объёс-

няющим переменным 2012 г. Оба варианта приемлемы, мы выбрали второй, так как в реальности именно с такой задачей сталкивается кредитор средних и малых фирм.

Для наглядности на рис. 4 мы приводим ROC-кривые для алгоритма случайного леса по балансированным выборкам и наборам регрессоров согласно популярности в работах других исследователей. Площадь под ROC-кривой примерно одинакова для всех отраслей и составляет около $\frac{3}{4}$.

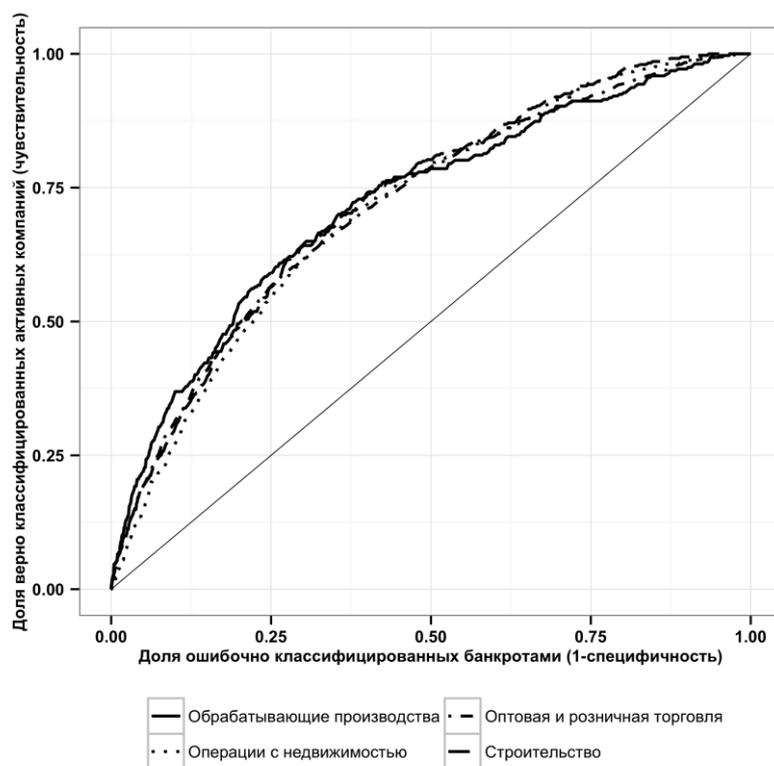


Рис. 4. ROC-кривая для алгоритма случайного леса по отраслям

6.2. Выбор алгоритма оценивания

Сравним результаты для различных методов и отраслей. Значения площади под ROC-кривой показаны на рис. 5. Чёрными горизонтальными линиями изображены медианы значений для каждого метода. Каждому методу соответствует несколько значений (от 3 до 6), так как мы сравнивали шесть наборов объясняющих переменных.

Для начала стоит отметить, что все методы прогнозируют достаточно хорошо, потому что площадь под ROC-кривой всегда выше 0,5. Более того, вне зависимости от отрасли наилучшим методом является алгоритм случайного леса. У него наибольшие медианные значения площади под ROC-кривой, все либо равны, либо близки к 0,7, что считается высоким показателем прогнозной силы. Во всех случаях не только медиана, но и весь диапазон значений площади для этого метода лежит выше максимальных значений для остальных методов. Это говорит о нелинейной и неаддитивной зависимости между вероятностью стать банкротом и объясняющими переменными.

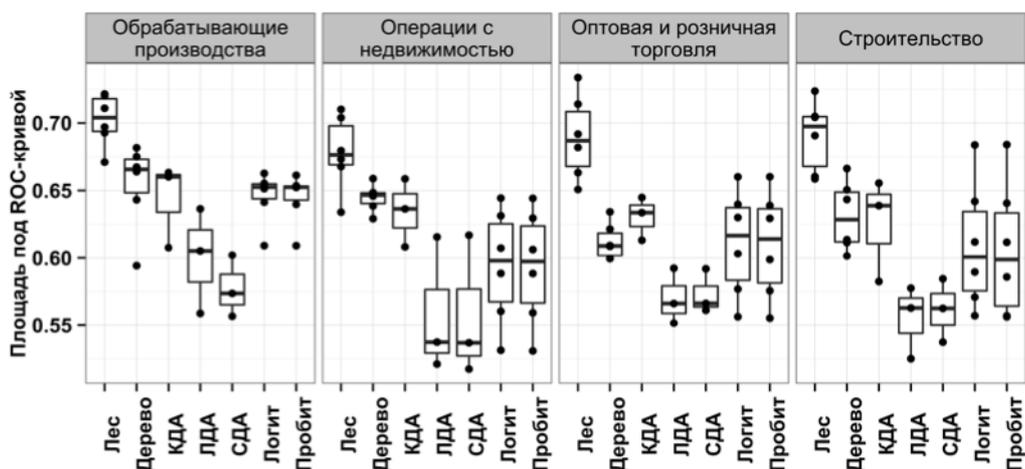


Рис. 5. Площадь под ROC-кривой для различных методов и отраслей

На втором месте оказывается либо классификационное дерево, либо КДА: это зависит от отрасли. Для обрабатывающих производств и оптовой и розничной торговли КДА превосходит дерево, а для двух оставшихся отраслей наоборот. Третьими идут логит- и пробит-модели: разброс значений площади под ROC-кривой для них наибольший во всех отраслях кроме обрабатывающих производств. ЛДА и СДА практически идентичны для каждой отрасли и идут на последнем месте.

На рис. 5 мы приводим результаты только для балансированных выборок, потому что оказалось, что превышение числа активных компаний в выборке по сравнению с неактивными не оказывает влияния на прогнозный потенциал того или иного метода. Например, для алгоритма случайного леса площадь под ROC-кривой по балансированным и небалансированным выборкам отличается на несколько процентов: медиана равна 2%, а среднее арифметическое – 3%.

6.3. Выбор набора регрессоров

В предыдущем разделе мы выяснили, что алгоритм случайного леса превосходит все остальные методы. По этой причине выбирать наилучший набор объясняющих переменных мы будем именно для этого метода. В табл. 6 приведены средние значения площади под ROC-кривой по балансированным и небалансированным выборкам для каждой отрасли и формулы в отдельности. Жирным шрифтом мы выделили максимальное значение для каждой отрасли.

Для трёх отраслей, кроме оптовой и розничной торговли, наиболее высокие значения площади под ROC-кривой соответствуют модели с расширенным набором нефинансовых характеристик и набором финансовых переменных, отобранных по частоте использования в исследованиях банкротств компаний. Однако именно для такого набора объясняющих переменных значения площади всегда выше 0,71 в отличие от других формул. Хотя набор финансовых переменных, отобранных с помощью LASSO, и со

всеми нефинансовыми показателями, и превосходит остальные наборы регрессоров для оптовой и розничной торговли, в двух отраслях те же финансовые переменные и только возраст компании дают наихудшие результаты. Поэтому наилучшим набором мы считаем набор популярных финансовых переменных и полный набор нефинансовых характеристик фирм.

Таблица 6. Значение площади под ROC-кривой для алгоритма случайного леса

Формула / отрасль	Обрабатывающие производства	Операции с недвижимостью	Оптовая и розничная торговля	Строитель- ство
LASSO + Возраст	0,646	0,634	0,671	0,666
LASSO + Все	0,705	0,666	0,723	0,715
Альтман + Возраст	0,688	0,667	0,644	0,653
Альтман + Все	0,709	0,707	0,695	0,698
Популярные + Возраст	0,701	0,675	0,657	0,682
Популярные + Все	0,728	0,714	0,714	0,725

Если попарно сравнивать формулы с одинаковыми финансовыми переменными и разным составом нефинансовых, то очевидно, что включение в модель всех нефинансовых показателей (размера компании, организационной формы, федерального округа), а не только возраста, значительно увеличивает прогнозную силу каждой модели. Таким образом, расширенные модели лучше базовых версий.

Также стоит обратить внимание, что здесь не видно значительной разницы между результатами для разных отраслей.

6.4. Влияние отдельных факторов на вероятность банкротства

Опишем зависимость вероятности банкротства от наилучшего набора объясняющих переменных (популярные финансовые отношения и все нефинансовые характеристики).

Для алгоритма случайного леса на рис. 6 приведём важность отдельных переменных по отраслям.

Важность конкретной переменной определяется (усреднённым по всем деревьям) суммарным падением индекса Джини по всем узлам, относящимся к данной переменной. Суммарное падение индекса Джини нельзя сравнивать напрямую между разными выборками. При большей выборке деревья, как правило, содержат большее количество узлов, и поэтому суммарное падение индекса Джини оказывается выше.

Качественный вывод из рис. 6 для всех отраслей одинаковый. Наименее важными нефинансовыми показателями оказываются организационная форма (*legal_form*) и размер компании (*lasize*). Наименее важным финансовым отношением оказывается показатель обслуживания долга *iptd* (*Interest paid / Total debt*). Важность остальных показателей примерно равна и существенно выше. Для строительства и операций

с недвижимостью на первое место по важности выходит показатель рентабельности *ebta* (EBIT / Total assets).

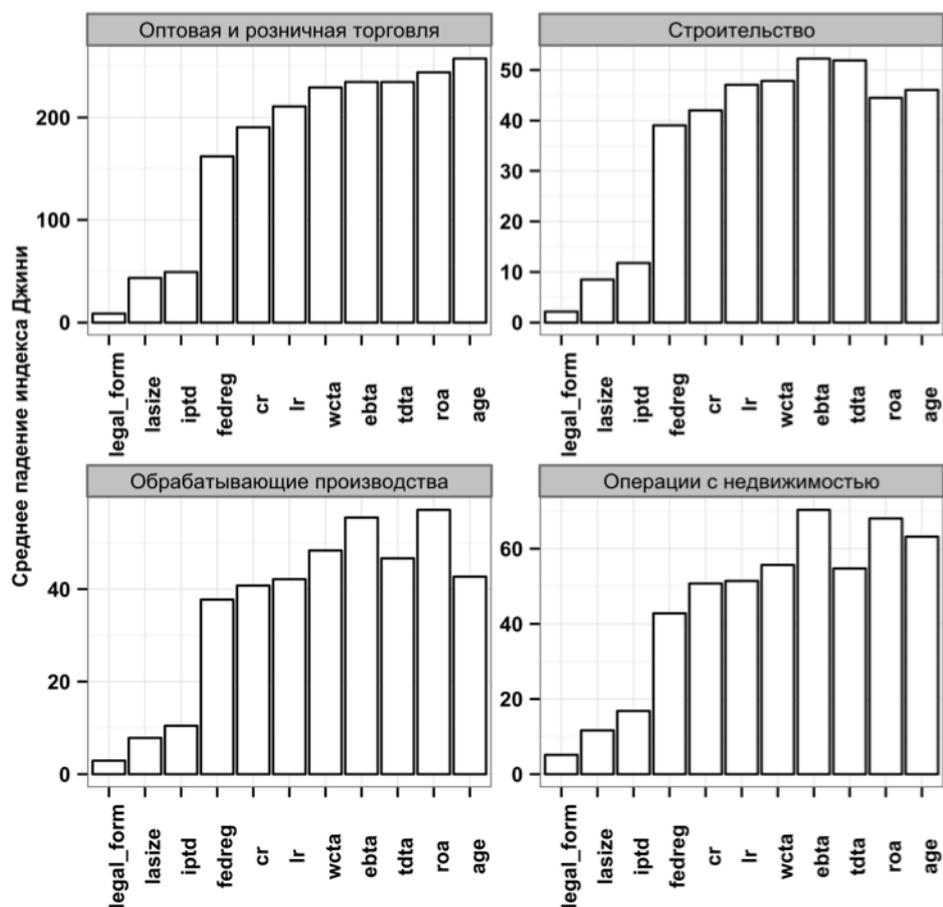


Рис. 6. Важность переменных для алгоритма случайного леса

К сожалению, визуализировать зависимость, обнаруженную алгоритмом случайного леса, крайне сложно. Случайный лес состоит из 500 деревьев, каждое дерево строится по своей подвыборке, при каждом делении узла используется свой набор объясняющих переменных. Мы наглядно демонстрируем направление связи между регрессорами и вероятностью дефолта с помощью двух подходов: предельных эффектов в логит-моделях и урезанных классификационных деревьев. Урезанное классификационное дерево существенно короче любого из тех деревьев, что строятся в алгоритме случайного леса, но его можно рассматривать как некое усреднённое дерево заданной высоты. Логит- и пробит-модели показывают одинаковые и стабильно хорошие результаты среди алгоритмов, оцениваемых методом максимального правдоподобия.

Предельные эффекты в логит-модели можно считать несколькими способами. В силу того, что распределение объясняющих переменных далеко от нормального и зачастую несимметрично, считать предельные эффекты для среднего значения регрессоров может быть ошибкой. Мы приводим предельные эффекты, усреднённые по всем

наблюдениям, для наилучшего набора регрессоров (популярные финансовые отношения плюс все нефинансовые характеристики) и балансированных выборок в табл. 7.

Таблица 7. Предельные эффекты в логит-модели по отраслям

	Обрабаты- вающие производства	Операции с недви- жимостью	Оптовая и розничная торговля	Строи- тельство
(Intercept)	-0,034 (0,041)	-0,186 (0,117)	0,058 (0,063)	-0,150 (0,110)
iptd	0,345 (0,090) ***	0,002 (0,002)	0,046 (0,025)'	1,579 (0,732) *
roa	-0,038 (0,006) ***	-0,002 (0,002)	-0,006 (0,003) *	-0,018 (0,012)
cr	-0,001 (0,001)	0,001 (0,001)	0,000 (0,000)	-0,002 (0,002)
wcta	-0,038 (0,012) **	0,000 (0,002)	0,021 (0,007) **	0,000 (0,008)
ebta	0,030 (0,006) ***	0,003 (0,003)	0,002 (0,003)	-0,004 (0,016)
lr	0,001 (0,001) 0,023	-0,001 (0,001)	0,000 (0,000)	0,005 (0,002) *
tdta	(0,007) ** -0,003	0,000 (0,002)	0,021 (0,007) **	0,006 (0,007)
age	(0,002) * -0,003	-0,008 (0,003) *	-0,010 (0,002) ***	-0,001 (0,004)
Far Eastern	0,042 (0,039)	0,252 (0,066) ***	0,159 (0,038) ***	0,164 (0,070) *
North Caucasian	0,073 (0,039)	0,024 (0,217)	0,168 (0,056) **	0,409 (0,156) **
Northwest	0,036 (0,024)	0,184 (0,046) ***	0,201 (0,025) ***	0,111 (0,050) *
Siberian	0,096 (0,040) *	0,151 (0,049) **	0,242 (0,021) ***	0,234 (0,045) ***
South	0,013 (0,025)	-0,117 (0,077)	0,086 (0,031) **	0,095 (0,057)
Ural	0,062 (0,030) *	0,068 (0,063)	0,095 (0,033) **	0,038 (0,062)
Volga	0,025 (0,021)	0,192 (0,042) ***	0,095 (0,024) ***	0,061 (0,045)
Lasize micro	0,006 (0,031) -0,072	0,257 (0,098) **	-0,013 (0,042)	0,067 (0,083)
Lasize small	(0,033) *	0,138 (0,104)	-0,027 (0,042)	-0,048 (0,083)
Limited liability company	-0,056 (0,027) *	-0,049 (0,051)	-0,114 (0,046) *	0,012 (0,073)
AIC	1 030,185	1 335,344	5 208,802	1 071,093
Num. obs.	800	1 002	3 892	806

***p < 0,001, **p < 0,01, *p < 0,05, 'p < 0,1.

В табл. 7 можно увидеть несколько интересных результатов. Во-первых, отрасли довольно сильно отличаются друг от друга. Например, в обрабатывающих производствах практически все финансовые переменные, кроме cg и lg , значимы при всех разумных уровнях значимости, а нефинансовые характеристики не так важны. Если же смотреть на результаты для отрасли операций с недвижимостью, то тут картина обратная: ни один предельный эффект для финансовых отношений не значим, а федеральные округа и размер компании оказывают значительное влияние. Оптовая и розничная торговля выделяется из всех отраслей тем, что все предельные эффекты для дамми на округа значимы при всех уровнях значимости, но в то же время некоторые предельные эффекты для финансовых переменных значимы. Для строительной отрасли результаты другие: только Сибирский федеральный округ значим при любом разумном уровне значимости, оставшиеся предельные эффекты значимы не всегда.

Для трёх отраслей, кроме строительства, предельный эффект возраста компании значим и отрицателен: вероятность дефолта падает с ростом возраста фирмы. Это логично, потому что давно существующие компании зачастую более устойчивы, чем только что открывшиеся бизнесы.

Показатель рентабельности goa значим для двух отраслей и отрицательно влияет на вероятность стать банкротом, т.е. чем выше рентабельность, тем ниже вероятность обанкротиться. Это соответствует логике, согласно которой если компания нерентабельна, она будет закрыта.

Показатель финансового рычага $tdta$ значим тоже для двух отраслей: с увеличением отношения суммарного долга к суммарным активам опасность банкротства растёт. Это можно объяснить тем, что величина накопленного долга по сравнению с величиной активов будет слишком велика, чтобы компания была в состоянии погасить его.

В дополнение к интерпретации предельных эффектов в логит-моделях проинтерпретируем урезанные классификационные деревья, представленные на следующей странице на рис. 7–10. Как и логит-модели эти деревья строились по балансированной выборке и оптимальному набору регрессоров. Деревья урезаны до четырёх терминальных узлов. Для всех отраслей вторым критерием отбора банкротов и активных предприятий выступает федеральный округ. Меньше всего банкротов в Центральном, Южном, Уральском и Приволжском федеральных округах. В качестве первого критерия для деления на банкротов и активные компании выступают показатели рентабельности: goa для оптовой и розничной торговли, $ebta$ для всех остальных отраслей. Возраст компаний является одним из критериев для двух отраслей: оптовой и розничной торговли и операций с недвижимостью. Возраст компании 6 и более лет является признаком активного предприятия. Таким образом, в деревьях представлены как финансовые, так и нефинансовые показатели, что подтверждает необходимость включения нефинансовых характеристик в модели.

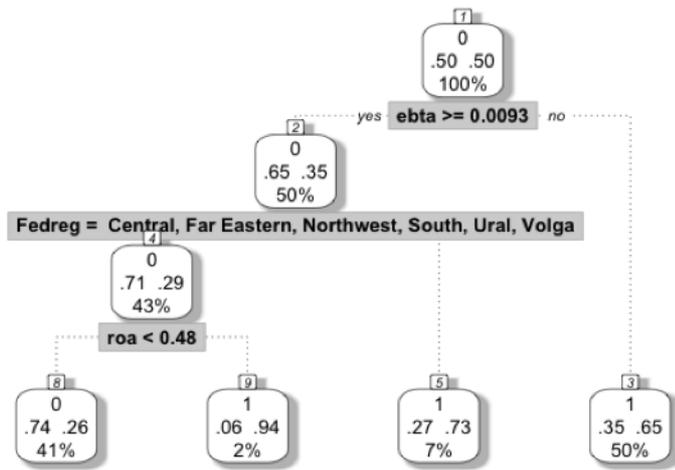


Рис. 7. Классификационное дерево для отрасли обрабатывающих производств

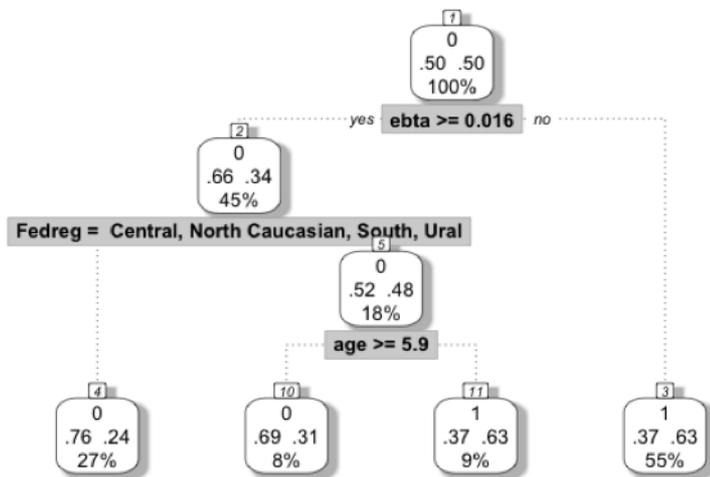


Рис. 8. Классификационное дерево для отрасли операций с недвижимостью

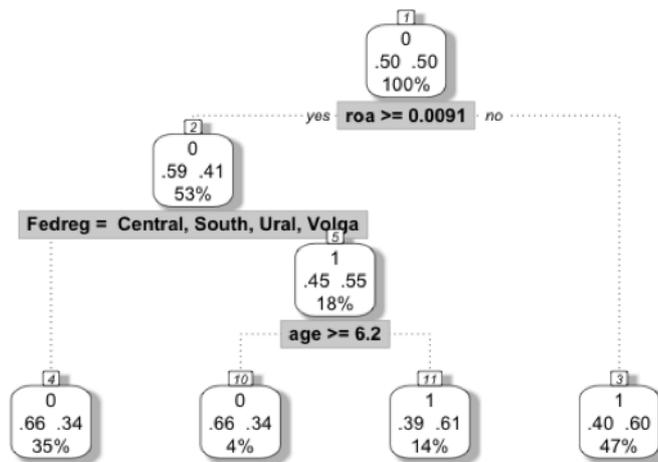


Рис. 9. Классификационное дерево для отрасли оптовой и розничной торговли

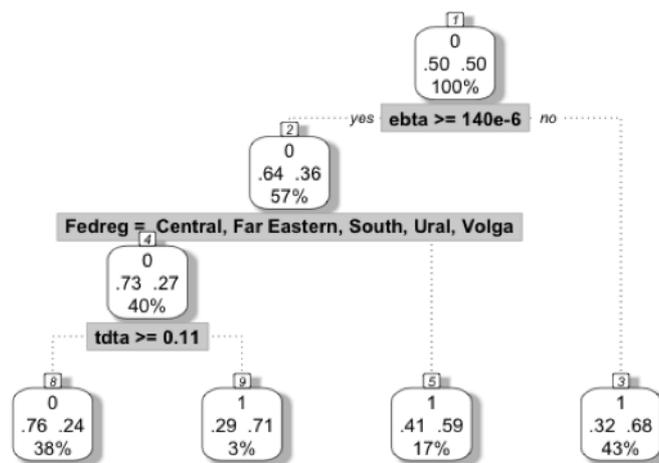


Рис. 10. Классификационное дерево для отрасли строительства

7. Выводы и перспективы исследования

Из отобранных нами для сравнения алгоритмов (ЛДА, КДА, СДА, логит- и пробит-модели, классификационное дерево, алгоритм случайного леса) наибольшую прогнозную силу показал алгоритм случайного леса вне зависимости от отрасли и типа выборки (балансированные и небалансированные). В целом существенно нелинейные алгоритмы показывают лучшие результаты. Поэтому нам представляется перспективным включение в сравнение других нелинейных алгоритмов: нейронных сетей, метода опорных векторов и бустинга. К сожалению, нелинейные алгоритмы, как правило, представляют собой «чёрный ящик», который позволяет достаточно хорошо прогнозировать, но не позволяет наглядно описать зависимость.

Важным выводом является сильное влияние некоторых нефинансовых показателей. Сильное влияние оказывают отрасль, федеральный округ и возраст предприятия. Менее важную роль имеет размер предприятия (существенно более низкое падение индекса Джини в алгоритме случайного леса, но иногда значимы дамми-переменные в логит- и пробит-моделях). Организационная форма оказалась практически несущественной (наименьшая важность согласно алгоритму случайного леса и незначимые коэффициенты в логит- и пробит-моделях).

Наилучшим набором объясняющих финансовых отношений оказался набор, построенный нами по популярности использования в других исследованиях. Среди финансовых отношений наиболее важными оказались коэффициенты рентабельности, финансового рычага и ликвидности.

В базе данных Руслана мы обнаружили существенное количество пропущенных наблюдений. На наш взгляд, решение проблемы пропущенных значений – это наиболее важный шаг дальнейшей работы.

Решать данную проблему можно по нескольким направлениям. Во-первых, можно попытаться агрегировать данные из других источников, например, базы данных Фира или Спарк. Здесь придётся столкнуться с тем, что финансовые показатели пред-

ставлены в разных формах (РСБУ и МСФО). Более того, наверняка возникнут расхождения даже в показателях, рассчитанных якобы по одной методологии. Во-вторых, можно применить различные алгоритмы заполнения пропусков. Наиболее перспективным нам представляется использование мер близости наблюдений из алгоритма случайного леса. В-третьих, само отсутствие наблюдения может быть важным признаком, значит, в качестве потенциальных регрессоров, возможно, имеет смысл взять индикаторы пропущенных наблюдений.

Заполнение пропусков позволит не только улучшить прогнозную силу алгоритмов, которые мы сравнили в данном исследовании, но и реализовать другие алгоритмы. Здесь интересно было бы реализовать алгоритмы, позволяющие оценивать важность и направление воздействия каждого регрессора в условиях сильной мультиколлинеарности. К таким методам, например, можно отнести LASSO, метод эластичной сети или байесовскую регрессию пик-плато (“spike and slab”).

Литература

Гиленко Е.В., Довженко С.Е., Федорова Е.А. (2012) Модели прогнозирования банкротства: особенности российских предприятий. ФГОБУВПО «Финансовый университет при Правительстве Российской Федерации». С. 85–92.

Жданов В.Ю., Афанасьева О.А. (2011) Модель диагностики риска банкротства предприятий авиационно-промышленного комплекса // Корпоративные финансы. № 4. С. 77–89.

Макеева Е.Ю., Бакурова А.О. (2006) Прогнозирование банкротства компаний нефтегазового сектора с использованием нейросетей // Общественные науки и современность. № 6. С. 22–30.

Altman E.I. (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy // The journal of finance. Vol. 23. No. 4. P. 589–609.

Altman E.I., Marco G., Varetto F. (1994) Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience) // Journal of Banking & Finance. Vol. 18. No. 3. P. 505–529.

Altman E.I., Sabato G. (2007) Modelling credit risk for SMEs: Evidence from the US market // Abacus. Vol. 43. No. 3. P. 332–357.

Beaver W.H. (1966) Financial ratios as predictors of failure // Journal of accounting research. Vol. 4. P. 71–111.

Bellovary J.L., Giacomino D.E., Akers M.D. (2007) A review of bankruptcy prediction studies: 1930 to present // Journal of Financial Education. Vol. 33. P. 1–42.

Berger A.N. (2006) Potential competitive effects of Basel II on banks in SME credit markets in the United States // Journal of Financial Services Research. Vol. 29. No. 1. P. 5–36.

Ciampi F., Vallini C., Gordini N. (2009) Using Artificial Neural Networks Analysis for Small Enterprise Default Prediction Modeling: Statistical Evidence from Italian Firms. Oxford Business & Economics Conference Proceedings, Association for Business and Economics Research (ABER). Vol. 1. P. 1–26.

Craig B.R., Jackson W.E., Thomson J.B. (2007) Does government intervention in the small-firm credit market help economic performance? Federal Reserve Bank of Cleveland.

Edmister R.O. (1972) An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction // *The Journal of Financial and Quantitative Analysis*. Vol. 7. No. 2. P. 1477–1493.

Falkenstein E., Boral A., Carty L. (2000) RiskCalc for private companies: Moody's default model. As published in *Global Credit Research*, May.

Härdle W.K. et al. (2007) The default risk of firms examined with smooth support vector machines. Discussion papers, German Institute for Economic Research. No. 757. P. 1–30.

Hunter J., Isachenkova N. (2001) On the Determinants of Industrial Firm Failure in the UK and Russia in the 1990s, ESRC Centre for Business Research, University of Cambridge.

James G. et al. (2013) *An introduction to statistical learning*, Springer.

Kaplinski O. (2008) Usefulness and credibility of scoring methods in construction industry // *Journal of Civil Engineering and Management*. Vol. 14. No. 1, 21–28.

Khorasgani A. (2011) Optimal accounting based default prediction model for the UK SMEs. Proceedings of ASBBS Annual Conference: Las Vegas. Vol. 18. No. 1.

Kolari J.W., Ou C., Shin G.H. (2006) Assessing the profitability and riskiness of small business lenders in the banking industry // *Journal of Entrepreneurial Finance*. JEF. Vol. 11. No. 2. P. 1–26.

Lugovskaya L. (2010) Predicting default of Russian SMEs on the basis of financial and non-financial variables // *Journal of Financial Services Marketing*. Vol. 14. No. 4. P. 301–313.

Makeeva E., Neretina E. (2013) The Prediction of Bankruptcy in a Construction Industry of Russian Federation // *Journal of Modern Accounting and Auditing*. Vol. 9. No. 2. P. 256–271.

Maleev V., Nikolenko T. (2010) Predicting Probability of Default of Russian Public Companies on the Basis of Financial and Market Variables.

Martin D. (1977) Early warning of bank failure: A logit regression approach // *Journal of Banking & Finance*. Vol. 1. No. 3. P. 249–276.

Ohlson J.A. (1980) Financial ratios and the probabilistic prediction of bankruptcy // *Journal of accounting research*. Vol. 18. No. 1. P. 109–131.

Pompe P.P., Bilderbeek J. (2005) The prediction of bankruptcy of small-and medium-sized industrial firms // *Journal of Business Venturing*. Vol. 20. No. 6. P. 847–868.

Sirirattanaphonkun W., Pattarathammas S. (2012) Default Prediction for Small- Medium Enterprises in Emerging Market: Evidence from Thailand // *Seoul Journal of Business*. Vol. 18. No. 2, 25–54.

Tam K.Y., Kiang M.Y. (1992) Managerial applications of neural networks: the case of bank failure predictions // *Management science*. Vol. 38. No. 7. P. 926–947.

Venables W.N., Smith D.M. (2002) *An introduction to R, Network Theory*.

Wei L., Li J., Chen Z. (2007) Credit risk evaluation using support vector machine with mixture of kernel. *Lecture Notes in Computational Science and Engineering*, 431–438.

Wilson R.L., Sharda R. (1994) Bankruptcy prediction using neural networks // *Decision support systems*. Vol. 11. No. 5. P. 545–557.

Zeitun R., Tian G., Keen K. (2007) Default probability for the Jordanian companies: A test of cash flow theory // *International Research Journal of Finance and Economics*. Vol. 8. P. 147–162.

Demeshchev, B., Tikhonova, A.

Default prediction for Russian companies: intersectoral comparison [Electronic resource] : Working paper WP2/2013/05 / B. Demeshchev, A. Tikhonova ; National Research University Higher School of Economics. – Electronic text data (2 Mb). – Moscow : Publishing House of the Higher School of Economics, 2014. – 27 p. – (Series WP2 “Quantitative Analysis of Russian Economy”).

The primary aim of this research is to compare diverse statistical models to predict critical financial state for Russian private small and medium-sized companies belonging to different sectors of economy.

We use the following methods: Linear Discriminant Analysis, Quadratic Discriminant Analysis, Mixture Discriminant Analysis, Logistic Regression, Probit Regression, Tree and Random Forest. Our dataset consists of approximately 1,000,000 observations from the Ruslana database and covers the period from 2011 to 2012.

Instead of standard definition of default we use the notion of critical financial state which means that we add companies liquidated as a result of legal bankruptcy to those liquidated voluntarily.

We study four industries in detail: construction, manufacturing, real estate activities, retail and wholesale trade. Comparing industries, we come up to several compelling conclusions. On the one hand, the difference between sectors is so significant that it cannot be overcome by including several dummy variables but by estimating separate models for each industry.

On the other hand, sectors are similar in several ways. Firstly, importance ranking of regressors is stable among sectors that are analysed. This results in unique optimal set of variables chosen out of six possible alternatives. To add, inclusion of non-financial characteristics improves predictive power greatly. While age of a company and federal region are the key non-financial variables, size of a company is less important, and legal form is the weakest predictor. Secondly, Random Forest outperformed other statistical approaches on all data sets. For this method area under ROC-curve (the applied comparison criterion) reaches up to $\frac{3}{4}$ which is the same for all industries.

This research will be of vital importance especially to banks and other credit organisations providing loans to small and medium businesses as well as to state regulators.

Key words: bankruptcy prediction; model comparison; small and medium enterprises; retail and wholesale trade; manufacturing, real estate activities; construction.

JEL classification: C14, C45, G30, G33.

Demeshchev Boris – senior lecturer, National Research University Higher School of Economics, Faculty of Economics, Department of Applied Economics; 119049, Moscow, Russian Federation, Street Shabolovka 28, room 112; E-mail: bdemeshev@hse.ru

Tikhonova Anna – first-year Master's student, programme “Financial economics”, ICEF, National Research University Higher School of Economics, Faculty of Economics, Department of Applied Economics; 119049, Moscow, Russian Federation, Street Shabolovka 28, room 112; E-mail: annette.tikhonova@gmail.com

Препринт WP2/2014/04
Серия WP2
Количественный анализ в экономике

Демешев Б. Б., Тихонова А. С.

**Прогнозирование банкротства российских компаний:
межотраслевое сравнение**