Национальный Исследовательский Университет
Высшая Школа Экономики

Факультет Экономики

Профиль специальных дисциплин
«Математические методы анализа экономики»
Кафедра математической экономики и эконометрики

БАКАЛАВРСКАЯ РАБОТА

«Прогнозирование банкротства средних и малых российских компаний»

Выполнила
Студентка группы № 41ММАЭ
Тихонова Анна Сергеевна

Научный руководитель
Старший преподаватель
Демешев Борис Борисович

Москва, 2014

Содержание

1 Вступление

Хотя концепция кредитоспособности сама по себе не нова, интерес к ней сильно повысился в последние годы после кризиса 2008 — 2009 годов. Тема банкротств предприятий привлекает внимание исследователей и становится всё более актуальной в последнее время также в связи с требованиями, предъявляемыми Базельскими соглашениями. До Базельского соглашения (2007) требования к капиталу крупных и малых компаний были более размыты, однако после этого соглашения различия стали настолько ощутимы, что у банков появились новые стимулы для построения различных моделей для фирм разного размера.

Цель данной работы — моделирование вероятности серьёзных финансовых трудностей средних и малых российских непубличных компаний с помощью финансовых и нефинансовых показателей. Для достижения этой цели необходимо выполнить следующие задачи: сравнить различные подходы к прогнозированию вероятности; выявить различия моделей до, в течение и после кризиса 2008 — 2009 годов; проверить гипотезу о влиянии нефинансовых показателей на вероятность банкротства; проверить гипотезу о различиях моделей по отраслям и правовым формам.

Исследования для компаний среднего и малого бизнеса необходимы, потому что они являются прочной основой для экономики страны и способствуют инновациям и развитию. В то же время построение скоринговых систем для компаний такого размера — одна из основных задач банков, которые составляют свой кредитный портфель. К тому же крупные компании, акции которых котируются на бирже, легче оценить, чем более мелкие компании. По этой причине прогнозирование банкротства

средних и малых предприятий необходимо банкам и иным кредитным организациям, принимающим решение о предоставлении кредита фирмам на основе финансовой отчётности.

Новизна работы проявляется в следующем. Впервые в одной работе сравнивается большое количество статистических методов таких, как логит- и пробит-модели, метод опорных векторов, метод классификационных деревьев, алгоритм случайного леса, линейный, квадратичный дискриминантный анализ и дискриминантный анализ смеси распределений. Впервые оценивается столь большой массив данных российских предприятий. В исходной выборке содержится около миллиона наблюдений. Также сравниваются модели до и после кризиса 2008 — 2009 годов, что не было сделано ранее. В первый раз по российским данным сделана попытка оценить и учесть неоднородность по отраслям и формам организации предприятия. Период наблюдения: 2004 — 2012 годы. Более того, понятие банкротства расширено до понятия закрытия из-за серьёзных финансовых сложностей, не совместимых с дальнейшим продолжением деятельности компании, то есть исследуются действующие на данный момент компании, а также два типа неактивных: ликвидированные в результате банкроства и добровольно ликвидированные компании. Важно отметить, что используется отчётность компаний, адаптированная к международным стандартам финансовой отчетности (МСФО).

2 Обзор литературы

2.1 Определение банкротства и размера

В современном мире различия в том, как функционируют компании разного размера, огромны. Например, размер предприятия накладывает ограничения на допустимый размер долга. Так, для компаний малого бизнеса гораздо сложнее взять кредит, чем для крупных известных корпораций. По мнению Колари, У и Шина (2006, стр. 13), маленькие компании гораздо более рискованны, чем крупные компании. Более того, многочисленные исследования такие, как работа Сирираттанафонкуна и Паттаратаммаса (2012, стр. 27), показали, что эти типы компаний должны изучаться в отдельности. Обычно все компании делят на две группы: компании крупного бизнеса и компании малого и среднего бизнеса. Компании малого и среднего бизнеса, в свою очередь, делятся на три подкатегории: микро, малые и средние предприятия.

Стоит отметить, что существует множество подходов к определению размера компании. Однако в данной работе в качестве критериев для разделения на разные категории взяты два параметра: численность персонала (number of employees) и выручка (turnover). Именно эти два критерия выбора категории указаны в Федеральном законе № 209-ФЗ «О развитии малого и среднего предпринимательства в Российской Федерации». В таблице ниже представлены требования к каждому показателю для компаний, чтобы относиться к той или иной категории. Также для сравнения приведены значения этих же параметров для европейских предприятий. Компания каждого размера должна удовлетворять обоим критериям одновременно.

Таблица 1: Деление компаний по размеру в России и Европейском союзе

Размер	**Численность персонала**		**Выручка**	
	Европейский союз	**Россия**	**Европейский союз**	**Россия**
Микро	1—10	1—15	$\leq$ €2 млн	$\leq$ €1.4 млн (RUB 60 млн)
Малое	11—50	16—100	$\leq$ €10 млн	$\leq$ €9.6 млн (RUB 400 млн)
Среднее	51—250	101—250	$\leq$ €50 млн	$\leq$ €24 млн (RUB 1 млрд)

В числе требований, которые накладывает 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, стр. 2). 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.

В современной литературе существует два основных определения дефолта. Часть исследователей (например, Малеев, Николенко, 2010, стр. 1) понимают под дефолтом невозможность выплаты платежей процентов или основного тела долга. Другая часть исследователей, которая более многочисленна, в качестве определения дефолта использует понятие легального банкротства и его различные стадии. Среди сторонников этого определения есть и российские, и иностранные авторы: Хантер и Исаченкова (2001, стр. 6), Валлини, Сиампи и Гордини (2009, стр. 7), Кхорасгани (2011, стр. 155). Это определение дефолта кажется более логичным, потому что неспособность платить по своим обязательствам в данный момент не означает невозможности погашения задолженности в предусмотренные законом сроки.

2.2 Методы

Существует три подхода к прогнозированию риска банкротства: параметрические и непараметрические модели и теории опционов. Первые два подхода являются статистическими, в то время как третий связан с правилом отсутствия арбитража.

Первые шаги в прогнозировании банкротства компаний были сделаны ещё в 1960-х годах. Бивер (1966) предложил использовать анализ относительных показателей (одномерный параметрический метод), и Альтман (1968) стал применять линейный дискриминантный анализ (ЛДА). Основной недостаток метода Бивера состоял в выборе порога отсечения. В то же время модель Альтмана (Z-score model), которая основана на заранее определённых финансовых показателях, не учитывает некоторые источники дохода различных компаний. Более того, некоторые исследования (например, Луговская 2009, стр. 312) показали, что наличие нефинансовых переменных в модели сильно влияет на результат.

Следующим этапом в развитии прогнозирования банкротств стало применение логит- и пробит-моделей Мартином (1977) и Олсоном (1980), которые показали, что эти методы зачастую превосходят дискриминантный анализ. К тому же Олсон предложил оценивать вероятность дефолта не только на год, но и на два года вперёд.

Позже широкое распространение в данной сфере приобрели метод опорных векторов, применённый, например, Хэрдлом, Ли, Шэфером и Йехом (2007), а также нейронны сети (Тэм и Кианг, 1992; Альтман, Марко и Варетто, 1994).

Однако все эти методы имеют недостатки. Линейный дискриминантный анализ основывается на предположении о линейной функциональ-

ной зависимости. Логит- и пробит-модели требуют добавления переменных, чтобы ввести немонотонную зависимость между дефолтом и объясняющими переменными. В то же время некоторые методы вызывают сомнения, потому что финансовые отношения, рассчитанные с помощью отчёта о прибылях и убытках, зачастую не имеют нормального распределения и не являются независимыми (например, Олсон, 1980, стр. 112; Уилсон и Шарда, 1994, стр. 546). Вей, Ли и Чен (2007, стр. 431) также подчёркивают, что ЛДА может неверно классифицировать исходы, так как ковариационные матрицы дефолтных и активных компаний, скорее всего, не идентичны, в то время как проблемы, присущие деревьям, — сверхподгонка и локальная оптимальность оценки.

2.3 Факторы, влияющие на вероятность банкротства

Существует множество финансовых показателей, которые рассчитывает каждая компания. По этой причине перед каждым исследователем стоит вопрос, какие из них выбирать. Однако чаще всего выделяют пять групп показателей: рентабельность, ликвидность, оборачиваемость, финансовый рычаг, обслуживание долга. Так, в работе Альтмана и Сабато (2007, стр. 15), 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).

Что касается нефинансовых характеристик, необходимо рассматривать не только размер компании, но и её возраст, организационную форму, отрасль, к которой она относится. Помпе и Бильдербик (2005, стр. 848) отмечают, что прогнозировать вероятность банкротства молодых фирм сложнее, чем давно существующих компаний. Они предлагают оценивать различные модели по возрастным категориям. Фалькенстейн, Борал и Карти (2000, стр. 1) отмечают, что связь между финансовыми показателями и риском дефолта различна для публичных и непубличных компаний. Цейтун, Тиан и Кин (2007, стр. 5), Каплински (2008, стр. 21) утверждают, что методы необходимо адаптировать в зависимости от отрасли.

Среди работ, в которых рассматривается прогнозирование дефолтов именно компании среднего и малого бизнеса, стоит назвать Эдминстера (1972), Альтмана и Сабато (2007), Валлини, Сиампи и Гордини (2009).

2.4 Прогнозирование банкроства компаний в России

В то же время работ по прогнозированию банкротства российских компаний довольно мало. В 1990-е годы это было связано с коренными переменами в регулировании бизнеса, законах и ведении бухгалтерского учёта. Однако середине 1990-х годов была работа по сравнению банкротств российских и британских фирм Хантера и Исаченковой (2001). К сожалению, из-за малого числа наблюдений обобщить их результаты невозможно. Позже попытки Зайцевой, Сайффулина и Кадикова адаптировать модель Альтмана (Z-score model) и модель Олсона (O-score model)

не оказались успешными.

Луговская (2009) анализирует российские дефолтные компании с помощью ЛДА, Жданов и Афанасьева (2011) используют ЛДА и логит-модель, Макеева и Бакурова (2012) применяют нейронные сети, а Фёдорова, Гиленко и Довженко (2013), как и Макеева и Неретина, концентрируются исключительно на логит- и пробит-моделях. Кроме Луговской все изучают одну отрасль, например, строительство, нефтегазовую или авиастроительную отрасль, и не учитывают размер компаний. К тому же размер выборки во многих случаях крайне мал. Выбор финансовых отношений ставится под вопрос, так как отчётность по российским стандартам отличается от отчётности по международным, поэтому неверно было основываться на иностранных источниках для выбора показателей. Более того, к сожалению, в большинстве работ правовые формы никак не обозначены.

3 Особенности работы

В данной работе исследуются непубличные российские компании среднего и малого бизнеса. Определение размера соответствует критериям, представленным в федеральном законе РФ № 209-ФЗ «О развитии малого и среднего предпринимательства в Российской Федерации». Два типа непубличных компаний — это общества с ограниченной ответственностью (ООО) и закрытые акционерные общества (ЗАО).

В России средний и малый бизнес в основном представлен компаниями этих правовых форм, в то время как более крупные компании в основном представлены открытыми акционерными обществами (ОАО), а для прогнозирования вероятности дефолта такого рода публичных компаний применяется другой подход.

Понятие легального банкротства расширено до понятия серьёзных финансовых сложностей. В анализ включены активные компании и компании, которые из-за финансовых сложностей были ликвидированы, то есть ликвидированные банкроты и ликвидированные добровольно. Это объединение сделано из-за того, что в соответствии с процедурой банкротства и ликвидации компаний оба типа сталкиваются с невозможностью продолжения работы, отличие лишь в одном — банкроты накопили столько долга, что уже не могут по нему расплатиться, в то время как добровольно ликвидированные компании закрываются, так как бизнес приносит только убытки, а продать его не получается. Однако важно отметить, что если компания хочет доброльно прекратить свою работу, но стоимость имущества не покрывает все обязательства, то компания будет ликвидированна по процедуре банкротства.

В таблице 2 представлено краткое описание процедур банкротства и добровольной ликвидации. Более подробно можно посмотреть в Федеральном законе № 127-ФЗ «О несостоятельности (о банкротстве)» и в Федеральном законе № 129-ФЗ «О государственной регистрации юридических лиц и индивидуальных предпринимателей».

Таблица 2: Процедура банкротства и процедура 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

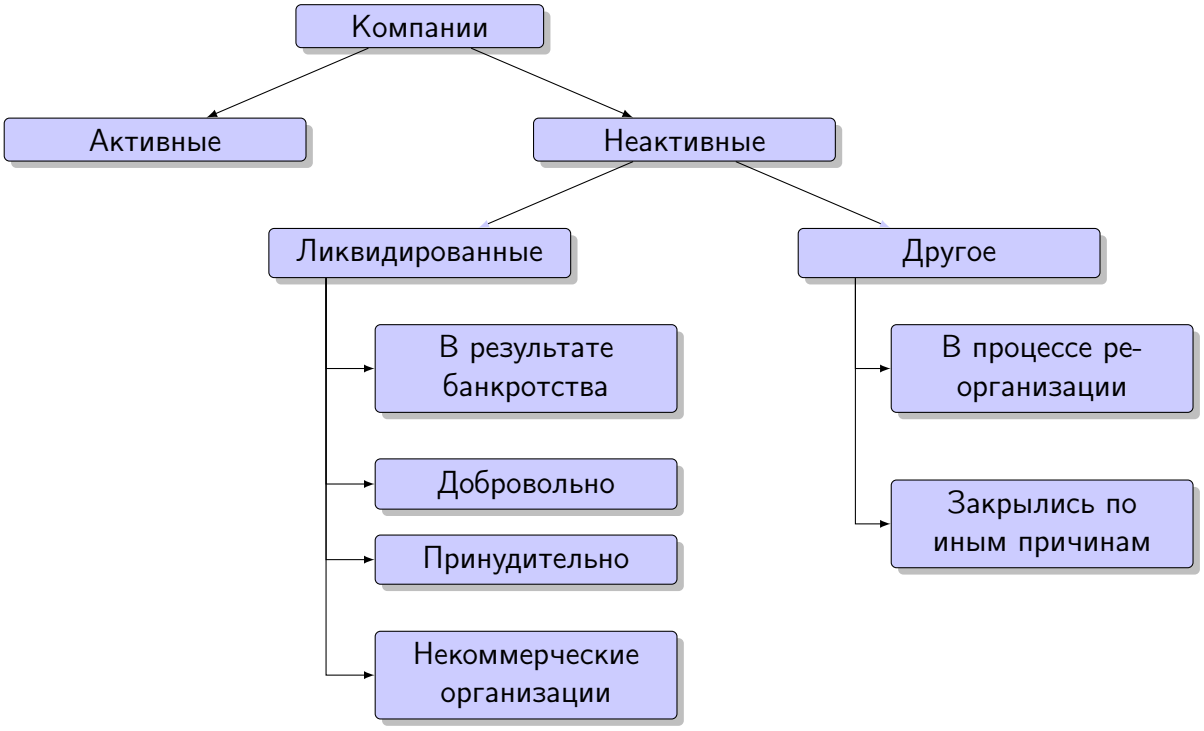
При прогнозировании банкротства используются как финансовые отношения, так и нефинансовые характеристики (возраст, отрасль, правовая форма, федеральный округ). Также взяты финансовые показатели, которые адаптированы к МСФО, имеют теоретическое обоснование.

4 Описание данных

Данные по российским фирмам собраны из базы данных российских, украинских и казахских компаний Руслана (`ruslana.bvdep.com`). Период исследования: 2004 — 2012 годы. Данные годовые. Среди исследуемых компаний содержатся общества с ограниченной ответственностью (ООО) и закрытые акционерные общества (ЗАО), которые относятся к одной из четырёх отраслей: строительству, обрабатывающим производствам, операциям с недвижимостью, оптовой и розничной торговле (по ОКВЭД).

Также исследуемые предприятия являются микро-компаниями, малыми или средними компаниями. Их можно разделить на три типа в зависимости от статуса: активные, добровольно ликвидированные и ликвидированные банкроты. На схеме ниже представлена классификация компаний по статусу.

Рис. 1: Деление российских компаний в зависимости от статуса



4.1 Нефинансовые переменные

Начнём с описания нефинансовых характеристик компаний. В таблице 3 представлены нефинансовые переменные, которые были изначально в данных или которые были созданы на основе имеющейся информации.

Таблица 3: Нефинансовые переменные в массиве данных

Переменная	Описание
Имеющиеся в данных	
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)
Созданные переменные	
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
ОКВЭД	6-значный код, укрупнённый до видов отраслей по ОКВЭД
Последний доступный размер	Микро, малое и среднее Размер на последний известный год, классифицированный с помощью Федерального закона № 209-ФЗ

Однако стоит описать создание некоторых переменных более подробно. Что касается индикатора дефолта (`def`), он определён следующим образом:

- Если компания прекратила своё существование до текущего года, она не включается в анализ (`def` = NA);

- Если компания ещё не была создана к текущему году, она не включается в анализ (`def` = 2);
- Если компания активна в текущем году, то она включается в анализ (`def` = 0);
- Если компания становится банкротом в текущем году, то она включается в анализ (`def` = 1);

Ещё одна нефинансовая переменная, создание которой необходимо описать, — возраст. Есть три возможные варианта:

1. Если компания активна (`def` = 0), возраст — разница между текущей датой и датой основания;
2. Если компания ликвидирована (`def` = 1), возраст — разница между датой присвоенного статуса и датой основания;
3. Если компания ещё не создана (`def` = 2), возраст равняется нулю.

Текущий год — каждый год из изучаемого периода от 2004 до 2012 в зависимости от того, на какой год строится модель, а текущая дата — поледний день текущего года (гггг-12-31).

Также необходимо пояснить, как создавалась переменная последний доступный размер, но для начала опишем сложности при введении переменной, обозначающей размер. Для каждого года берутся количество персонала компании и величина выручки. С их помощью производится деление в соответствии с законом Российской Федерации, но есть некоторые уточнения:

- Если данные по обоим критериям попадают в одну группу, то берётся размер этой группы;

- Если данные попадают в разные группы, то численность персонала выбирается основным критерием для выбора типа размера;
- Если есть данные только по одному из критериев, то размер определяется только относительно этого критерия;
- Если ни по одному из критериев нет данных, то ставится NA.

Однако даже при таком делении пропущенных значений в переменной размер осталось довольно много. Это вполне логично, потому что обычно в год дефолта компания не предоставляла данными ни по одному из необходимых критериев. С целью решения этой проблемы была введена другая переменная — последний доступный размер. Это означает, что если не было данных о размере на текущий год, то брался размер в предыдущем году, если его тоже не было, то брался размер два года назад и так далее.

4.2 Финансовые переменные

Что касается финансовых переменных, то из базы данных были взяты финансовые показатели из баланса и отчёта о прибылях и убытках, на основе которых были рассчитаны финансовые отношения. Разные авторы относят показатели в разные группы, но в данной работе классификация содержит пять групп: финансовый рычаг, ликвидность, рентабельность, обслуживание долга и активность. Эти группы соответствуют группам в работе Альтмана и Сабато (2007). Охарактеризуем каждую группу в отдельности.

Показатели финансового рычага характеризуют то, как соотносятся заёмные и собственные средства компании. Они отражают финан-

совую устойчивость фирмы и её степень риска. Также эти отношения могут отражать эффект увеличения прибыли за счёт взятия дополнительного долга из-за недостатка собственного капитала.

Показатели ликвидности характеризуют то, насколько быстро компания может превратить имеющиеся материальные ценности в деньги для покрытия своих финансовых обязательств. То есть чем больше у фирмы неликвидных активов, тем более высока вероятность того, что она в случае сложностей не сможет вовремя расплатиться по своим долгам.

Показатели рентабельности характеризуют степень эффективности использования средств компании. Они отражают, покрывает ли компания свои затраты доходами, если да, то получает ли компания прибыль.

Показатели обслуживания долга характеризуют кредитоспособность предприятия, отражая то, как быстро она совершает выплаты по долгу и какая доля денежного потока уходит на выплату процентов или основного тела долга.

Показатели активности характеризуют уровень деловой активности предприятия, то есть оборачиваемость средств компании. Они помогают компании определить необходимый уровень оборотных средств, потому что при ускорении оборачиваемости потребность в оборотном капитале снижается.

В приложении приведены таблицы отношений, входящих в каждую группу, а также расшифровка некоторых показателей. Деление переменных по группам проводилось с помощью прочитанных статей и собственного анализа.

5 Графический анализ данных

5.1 Нефинансовые переменные

Для начала необходимо привести число обанкротившихся компаний. В таблице 4 представлены доли (в процентах) этих компаний от числа всех предприятий четырёх исследуемых отраслей в каждом году.

Таблица 4: Доля банкротов от всех компаний (%)

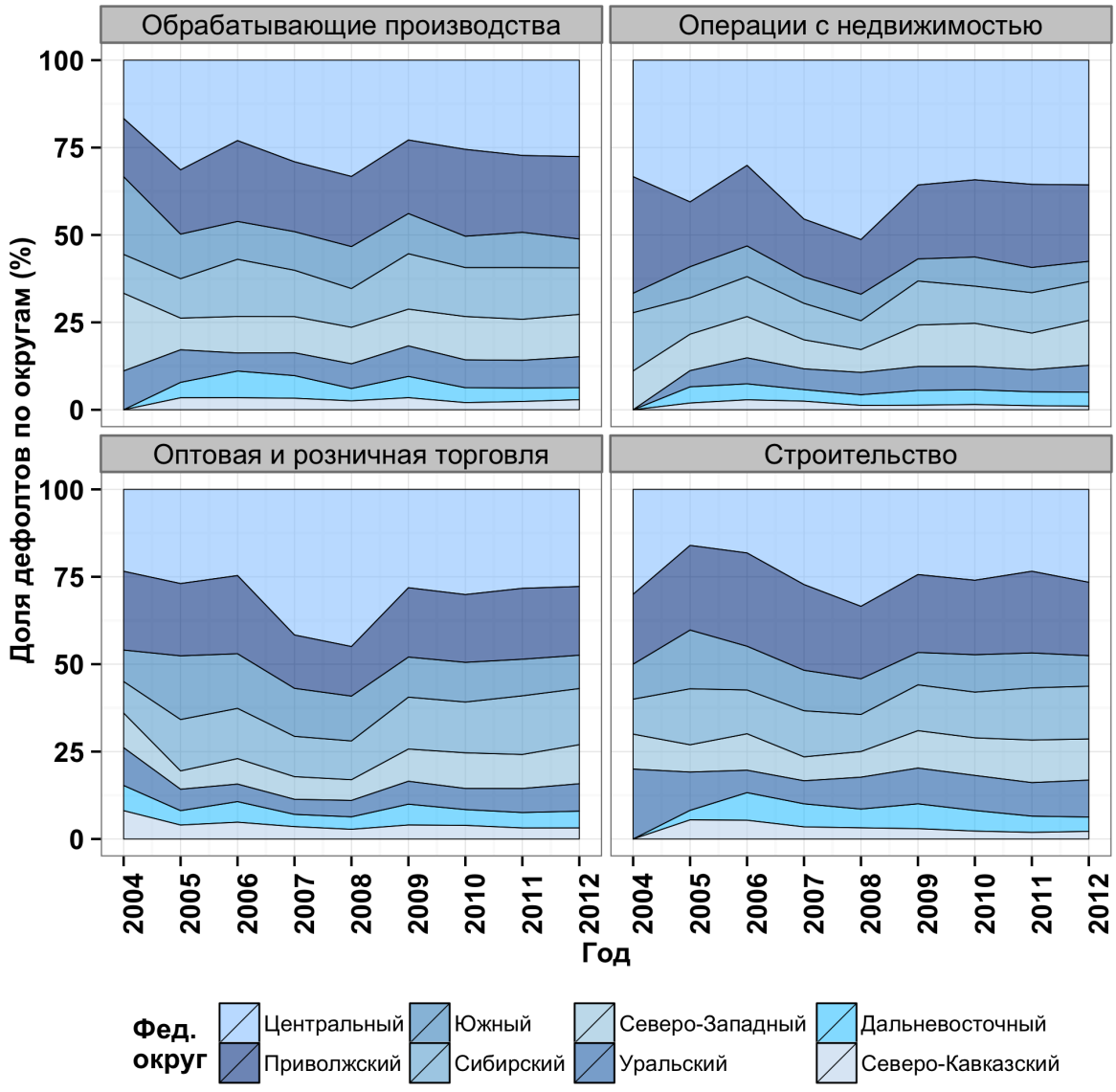
Год	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Доля	0.049	0.569	0.838	1.409	1.657	1.217	1.385	1.369	1.308

Несмотря на что доли кажутся маленькими, в абсолютном выражении численность банкротов за один год может достигать пятнадцати тысяч. Стоит отметить, что в 2008 году доля дефолтных компаний достигла своего максимума, потом начала падать, но всё-таки почти всегда оставалась выше докризисного уровня.

Более того, распределение дефолтов по федеральным округам различается в зависимости от отрасли, в которой работает компания. Оно представлено на графике 2. Лидерами по банкротствам стабильно были Центральный и Приволжский федеральные округа: их доли от всех дефолтов составляли 25% и больше. Однако в отрасли, связанной с недвижимостью, и оптовой и розничной торговле в 2008 году фирмы-банкроты в Центральном федеральном округе составляли около или ровно 50% всех дефолтов. Такое увеличение можно объяснить тем, что контракты на поставки подписываются на более короткий срок, чем контракты по строительству или обработке. Растущая неуверенность покупателей и отказ от контрактов жителей центральных районов России, где больше поставщиков услуг, чем в других округах, привёл к значительному росту доли дефолтов.

Увеличение дефолтов в целом в этот период неудивительно: мировой кризис наиболее сильно ударил по российской экономике именно в 2008 году. К тому же сложившаяся на тот момент обстановка, когда правительство основными целями объявило борьбу с ростом плохих кредитов и борьбу с инфляцией, полагая, что именно её увеличение дестабилизирует экономику, привела к росту ставки рефинансирования до 13%.

Рис. 2: Распределение дефолтных компаний по федеральным округам

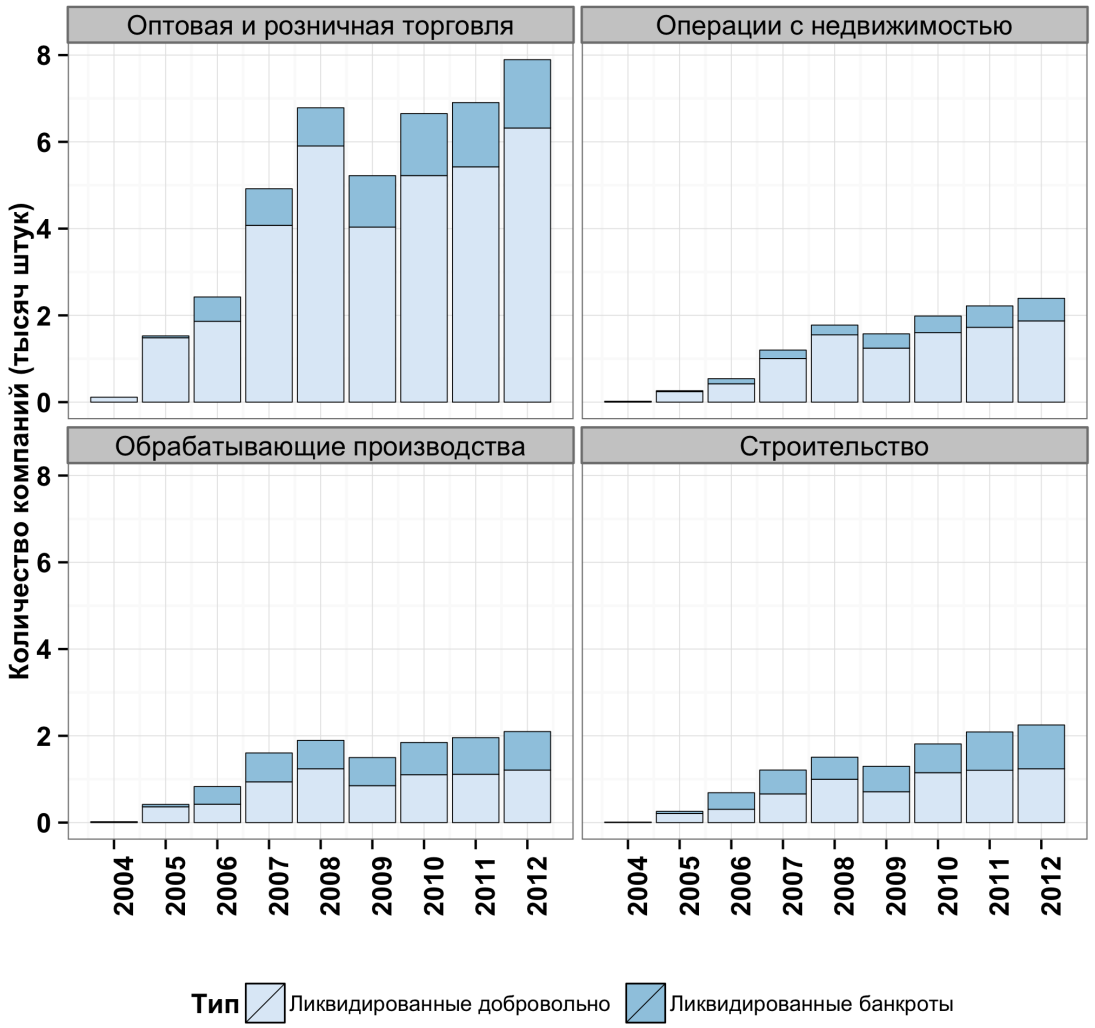


Вместе с повышением избирательности банков при выдаче кредитов и прекращением товарного кредитования это означало, что компании перестали иметь доступ к кредиту. Кризис кредитования на российском

рынке оказался даже сильнее европейского. В середине 2008 года необходимость государственной поддержки частного бизнеса стала для всех очевидна. После вливаний в экономику в 2009 году доля Центрального округа вернулась на докризисный уровень. Однако не все округа находятся под сильным влиянием ситуации во всей экономике. Так, в Дальневосточном и Северо-Кавказском округах банкротов было стабильно мало вне зависимости от отрасли. Более того, в 2004 году в этих округах в трёх из четырёх исследуемых отраслей совсем не было дефолтов.

Теперь рассмотрим компании по типу дефолта.

Рис. 3: Распределение дефолтных компаний по типу дефолта



Среди анализируемых отраслей больше всего дефолтов было в оптовой и розничной торговле. Причём в этой отрасли, как и в отрасли операций с недвижимостью число ликвидированных добровольно предприятий значительно больше, чем ликвидированных в результате банкротства. В оставшихся отраслях численность фирм в каждой из этих групп практически одинакова.

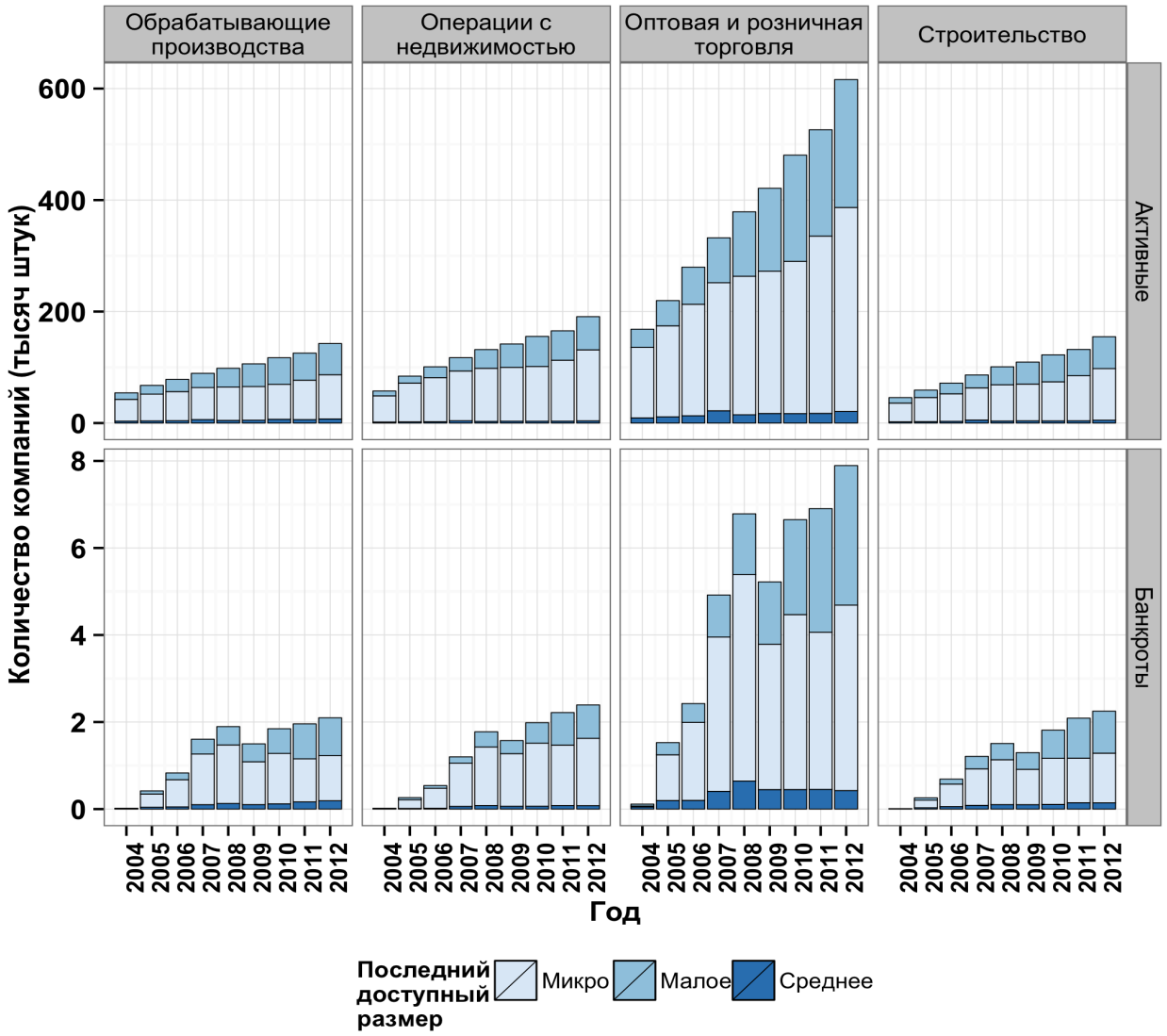
Такую значительную разницу в численности групп в оптовой и розничной торговле можно объяснить тем, что эти компании настолько малы, что зачастую не могут претендовать на большие суммы кредита в принципе, поэтому они в большинстве закрываются из-за невозможности приносить доход владельцу.

Более того, на графике 3 видно, что государственная поддержка в 2009 году помогла значительно уменьшить число банкротств. Однако стоит отметить, что после 2009 года количество компаний-банкротов вернулось к кризисному уровню и даже стало больше в 2011 и 2012 годах, причём это произошло за счёт прироста ликвидированных добровольно компаний. Это можно объяснить тем, что в посткризисные годы в России многим мелким компаниям стало нерентабельно функционировать.

В пользу этой мысли также выступает следующий график 4. На нём изображено количество активных компаний и компаний, прекративших своё существование из-за серьёзных финансовых сложностей. Компании также разделены по отраслям и по последнему доступному размеру. По графику сразу очевидно, что число банкротов росло одновременно с увеличением числа новых только что открывшихся компаний. Количество неактивных компаний достигало максимум восьми тысяч компаний в отрасли оптовой и розничной торговли. В то же время активных компаний

было всегда в разы больше, чем обанкротившихся предприятий.

Рис. 4: Распределение компаний по размеру и отраслям



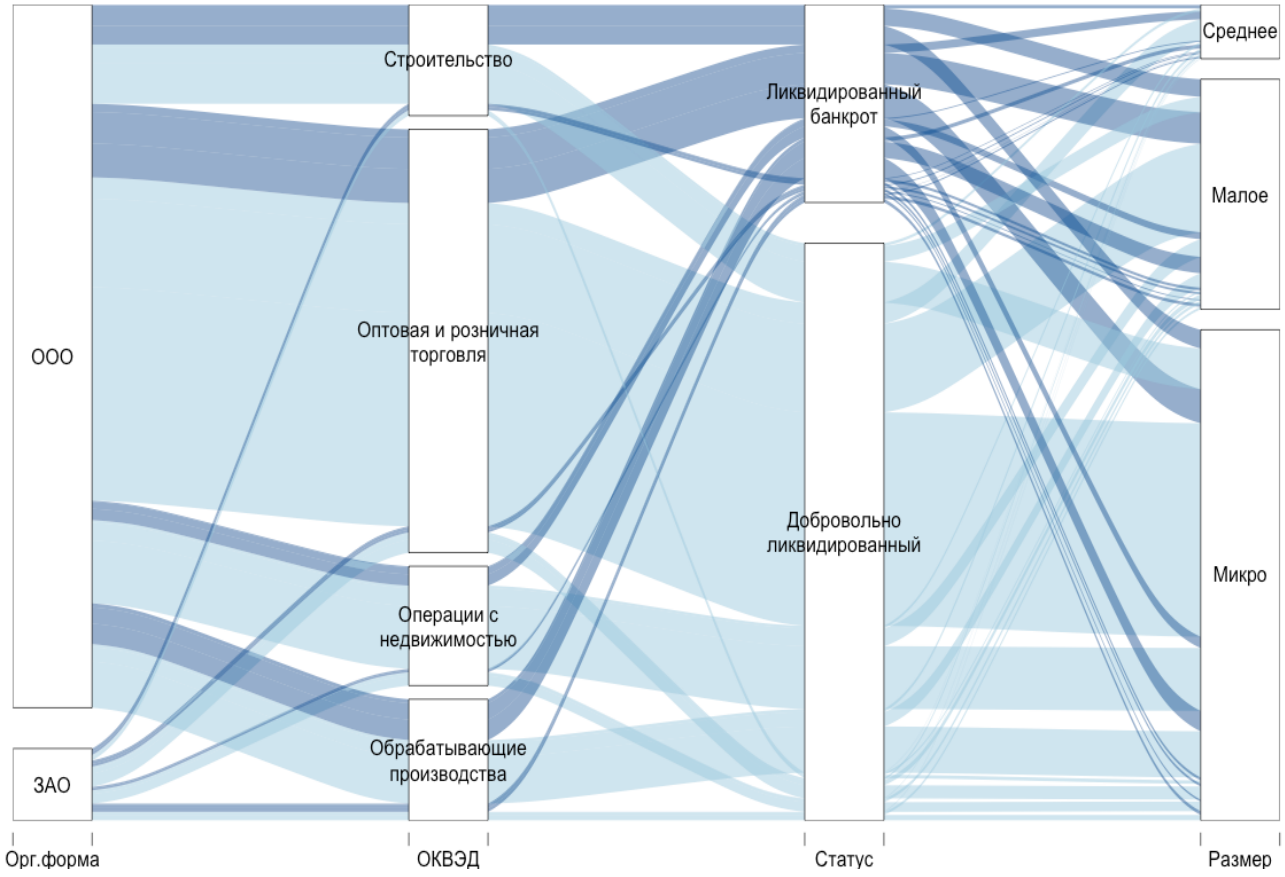
Оптовая и розничная торговля во все годы исследуемого периода насчитывала наибольшее число фирм, как среди функционирующих компаний, так и среди закрывшихся компаний. Остальные же отрасли более схожи друг с другом по численности предприятий в каждой группе.

Стоит обратить внимание на распределение компаний по размеру. Наиболее многочисленной группой оказались микро-компании, они представляют половину или более компаний в каждой из отраслей. На втором

месте идут компании среднего бизнеса. Средние по размеру фирмы представляют собой наименее многочисленную группу. Важно отметить, что такое распределение по группам характерно как и для активных компаний, так и для банкротов.

Теперь рассмотрим деление компаний по организационным типам за весь исследуемый период. На графике 5 тёмно-синим цветом выделены компании, ликвидированные в результате банкротства, а светло-синим — добровольно ликвидированные фирмы.

Рис. 5: Характеристики дефолтных компаний



Компаний, которые являются обществами с ограниченной ответственностью (ООО), гораздо больше, чем компаний, которые являются закрытыми акционерными обществами (ЗАО).

Это можно обосновать тем, что в ЗАО есть требования по выпуску акций, и тем, что ООО может быть более привлекательной организаци-

онной формой из-за более закрытого характера отношений между участниками. К тому же причиной может служить специфика анализируемых отраслей. Также в имеющихся данных ООО в основном имеют размер микро или малый. Более того, больше всего ликвидированных банкротов было в отрасли оптовой и розничной торговли. Они были обществами с ограниченной ответственностью. Далее следует строительный сектор, причём число ликвидированных добровольно равно числу ликвидированных из-за банкротства. График показывает, что по данным характеристикам нет серьёзных различий между двумя типами финансовых сложностей.

5.2 Финансовые переменные

Теперь посмотрим на финансовые переменные. В связи с тем, что в данных большое число финансовых отношений, в данном разделе будут представлены некоторые из них. Для целостности описания будут приведены все типичные законы распределения.

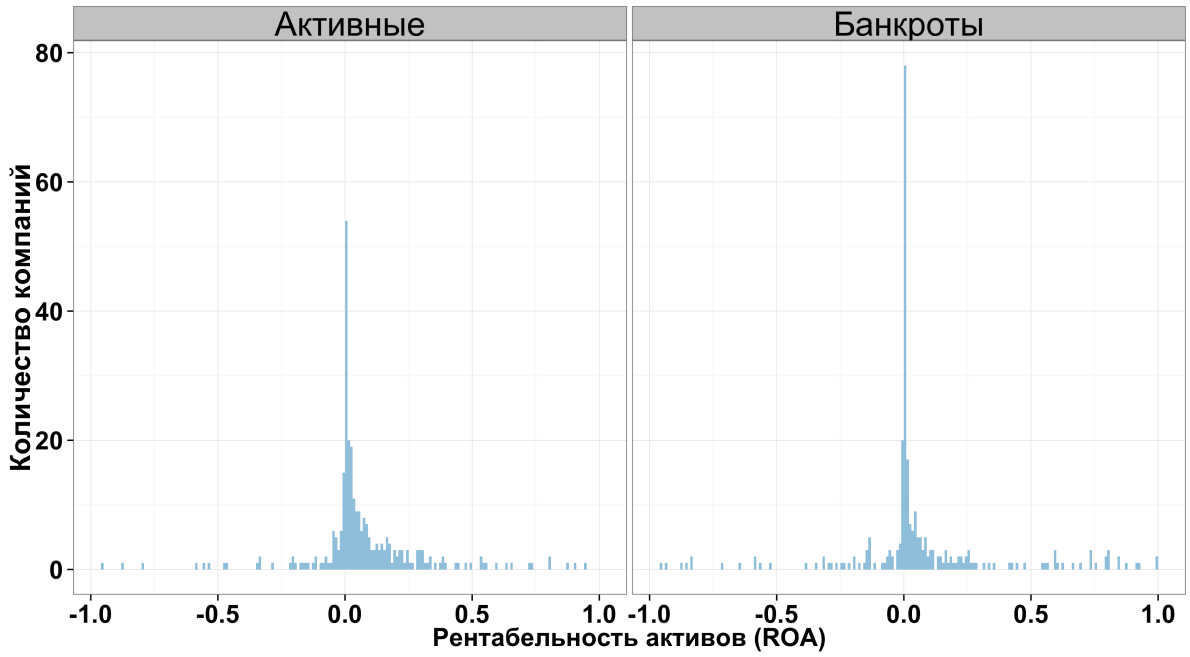
Более того, важно отметить, что практически все обанкротившиеся компании не подавали данных о финансовых показателях в год банкротства. Это делает невозможным иллюстрацию того, как отличаются показатели у компаний-банкротов на определённый год и активных в тот же год компаний. Поэтому для финансовых показателей будут приведены данные по финансовым переменным на текущий год и индикатором дефолта на следующий год, точно так же впоследствии будут строиться модели.

Ввиду того, что период анализа — 9 лет, то построим графики распределения нескольких переменных для одного года, а именно для 2012 года

как последнего имеющегося в данных. Для графического изображения была взята балансированная выборка (одинаковое количество активных компаний и компаний-банкротов) компаний оптовой и розничной торговли. Перед балансировкой данные 2012 года были очищены от пропущенных значений по изображаемым переменным.

На следующих двух страницах приведены графики четырёх показателей: показателя рентабельности, показателя ликвидности и двух показателей обслуживания долга. Показатель рентабельности (ROA) рассчитан как отношение Net income / Total assets, а показатель ликвидности равен (Current assets - stocks) / Current liabilities. Показатели обслуживания долга рассчитаны как Interest paid / Total debt, а также Turnover / (Total equity + non-current liabilities).

Рис. 6: Распределение рентабельности



Финансовые отношения не имеют нормального распределения. Более того, у них тяжёлые хвосты. Это заметно, даже несмотря на то что фактический диапазон значений этих переменных шире, чем представленный на графике.

Рис. 7: Распределение ликвидности

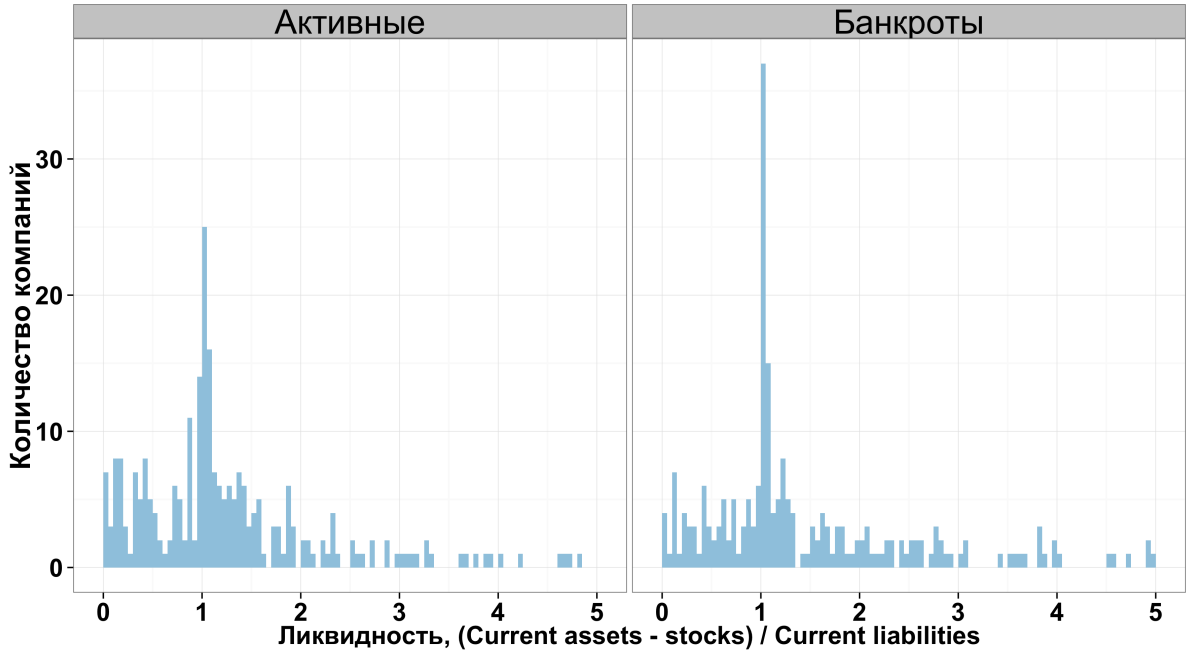
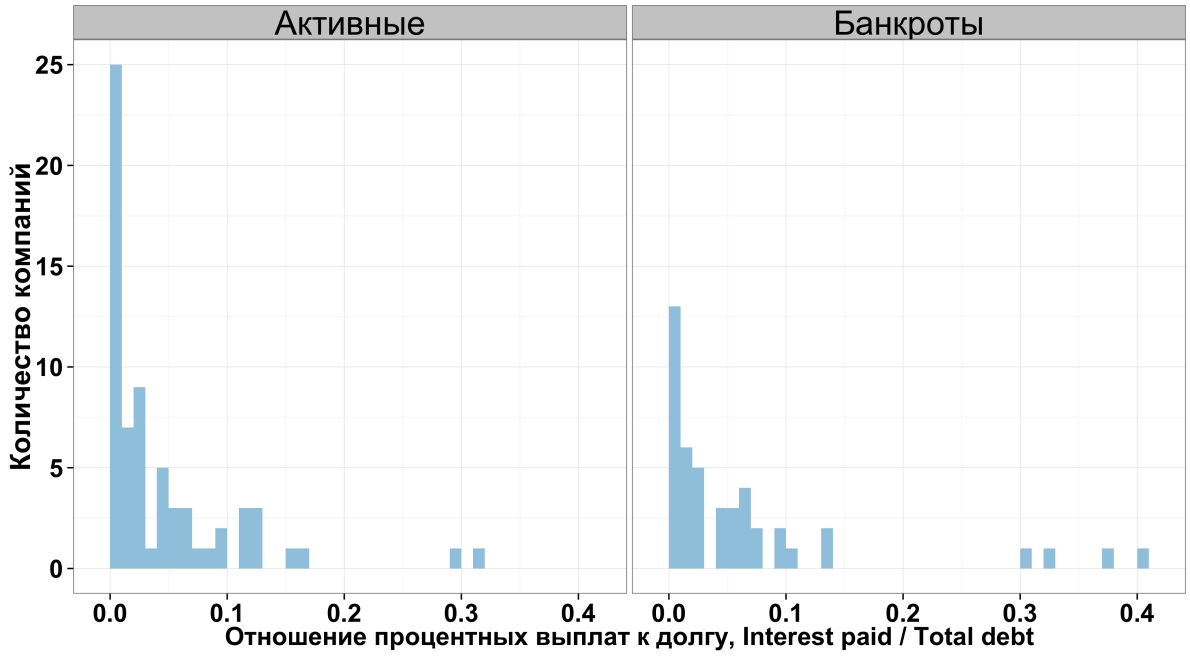
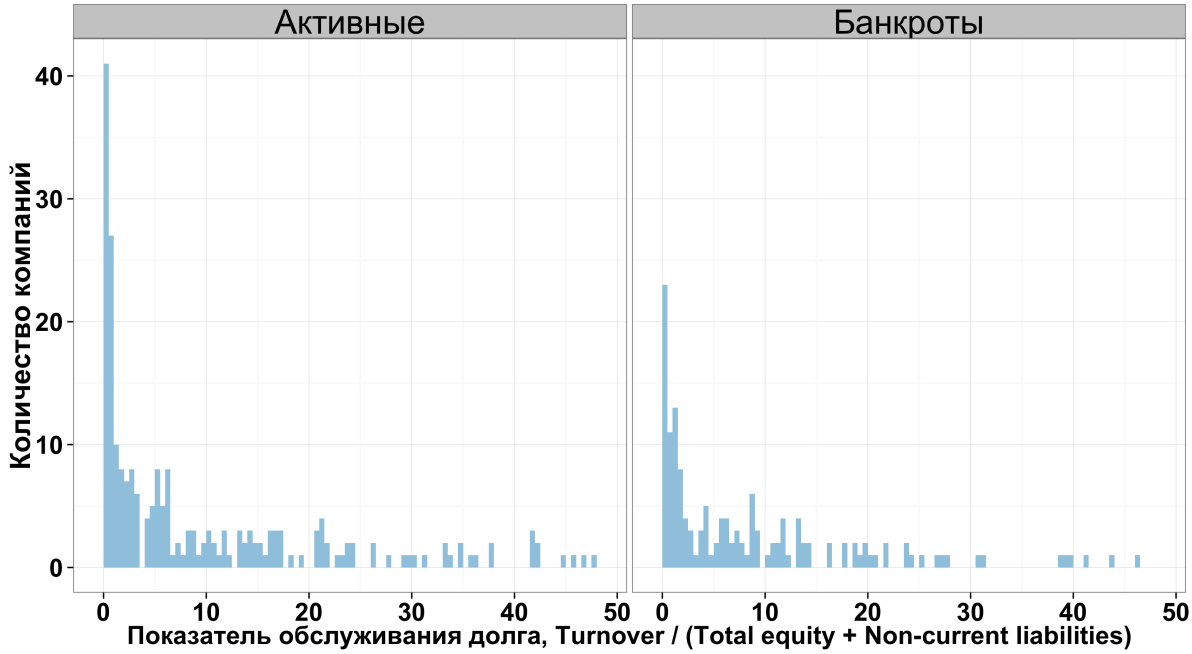


Рис. 8: Распределение отношения процентных выплат к долгу



Если сравнивать банкротов и активные компании, то распределения в целом похожи, небольшая разница, если и наблюдается, то не состоит в простом сдвиге распределения. Это говорит в пользу того, что эффект каждой переменной нелинейный и возможно следует учитывать взаимодействие переменных.

Рис. 9: Распределение показателя обслуживания долга



Далее представлены графики, отражающие медианы (median) и медианное абсолютное отклонение (Median absolute deviation, MAD) для четырёх групп финансовых переменных, в каждой из которой присутствуют по три представителя из каждой группы.

Медианное абсолютное отклонение рассчитывается по формуле:

$$MAD = 1.4826 \cdot \text{median}_i \left| x_i - \text{median}_j(x_j) \right|, \tag{1}$$

где median_i обозначает выборочную медиану. Домножение на константу 1.4826 используется для того, чтобы у нормального распределения медианное абсолютное отклонение равнялось стандартному отклонению.

Сначала стоит объяснить, почему используются медиана и медианное абсолютное отклонение, а не среднее значение и стандартное отклонение. Как было отмечено ранее, распределения финансовых переменных имеют очень тяжёлые хвосты, по этой причине среднее значение переменных очень сильно меняется и не может служить хорошим показателем изменения средней тенденции. И по этой причине стандартное отклонение

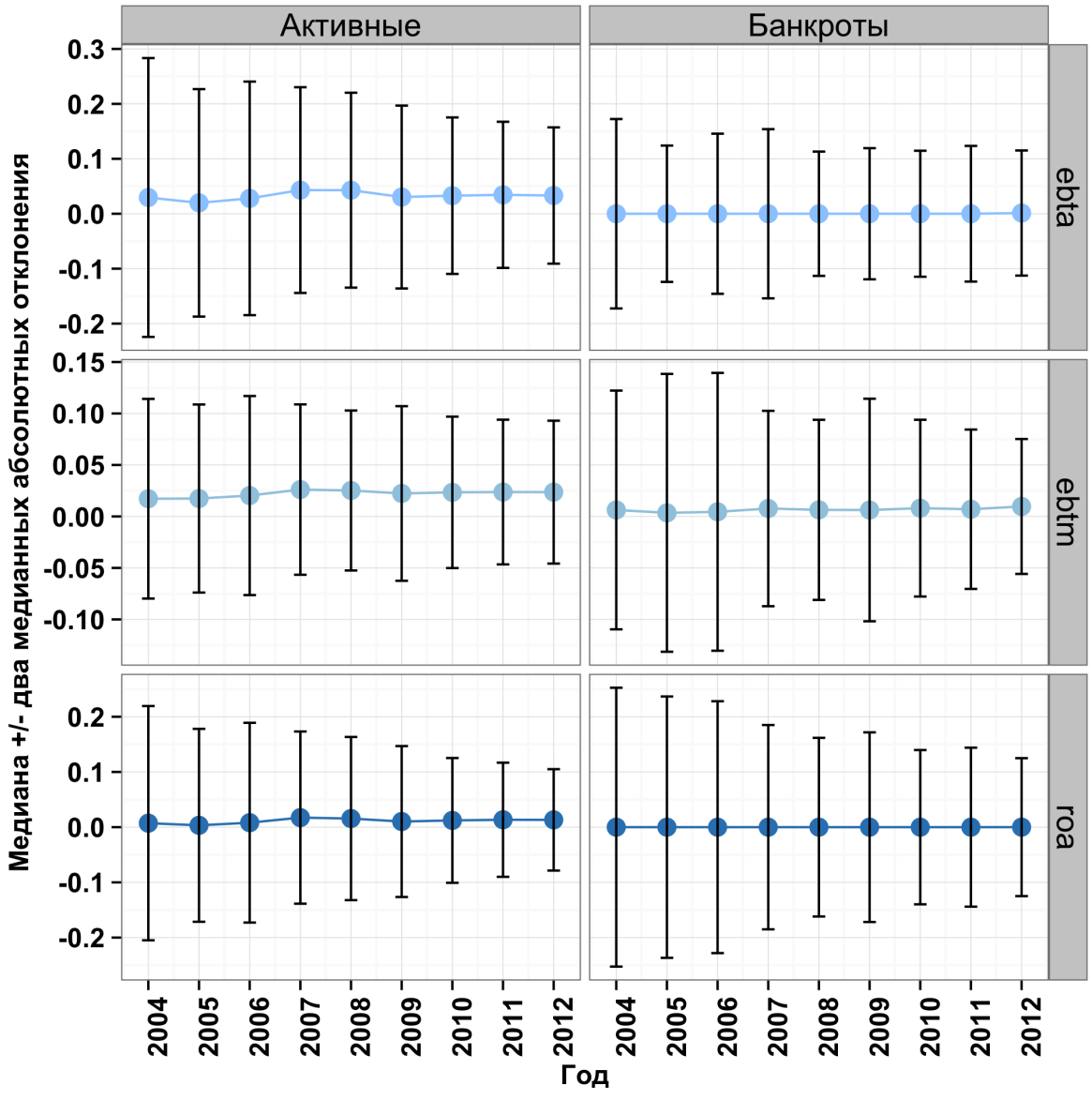
очень велико. Продемонстрируем это на следующем ниже примере.

Таблица 5: Характеристики переменной ROA для активных компаний

Год	Среднее значение	Медиана	Стандартное отклонение	Медианное абсолютное отклонение
2004	-0.589	0.007	130.570	0.106
2006	-0.331	0.008	26.691	0.091

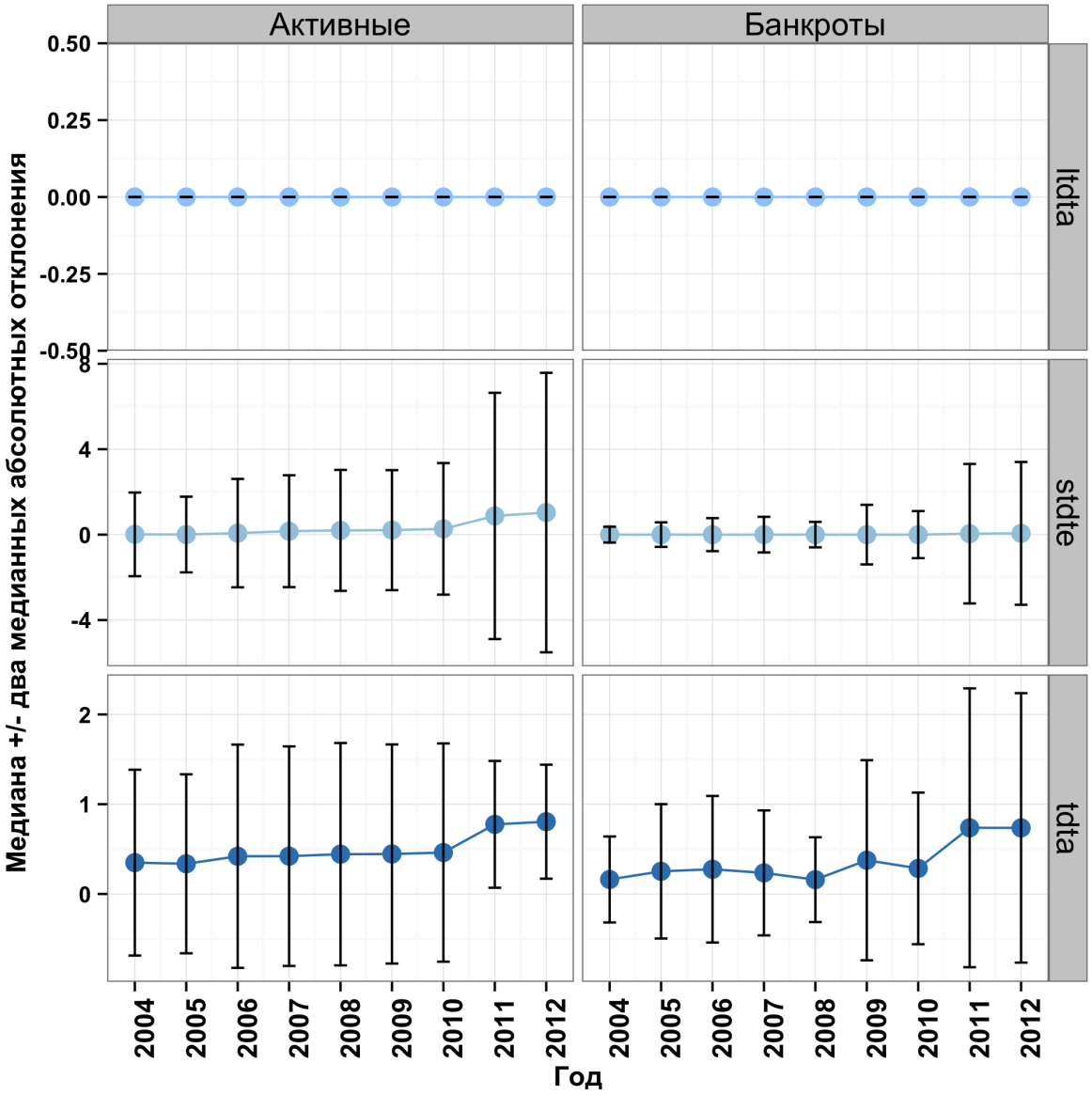
В таблице 5 приведены характеристики переменной рентабельности для активных компаний в 2004 и 2006 году. Стандартные отклонения для двух лет отличаются в разы.

Рис. 10: Показатели рентабельности



Теперь поясним, почему встречается четыре группы показателей, а именно показатели рентабельности, финансового рычага, ликвидности и обсулиживания долга. Пятая группа (показатели активности), заявленная в описании переменных, наименее распространённая, поэтому будут приведены четыре группы.

Рис. 11: Показатели финансового рычага

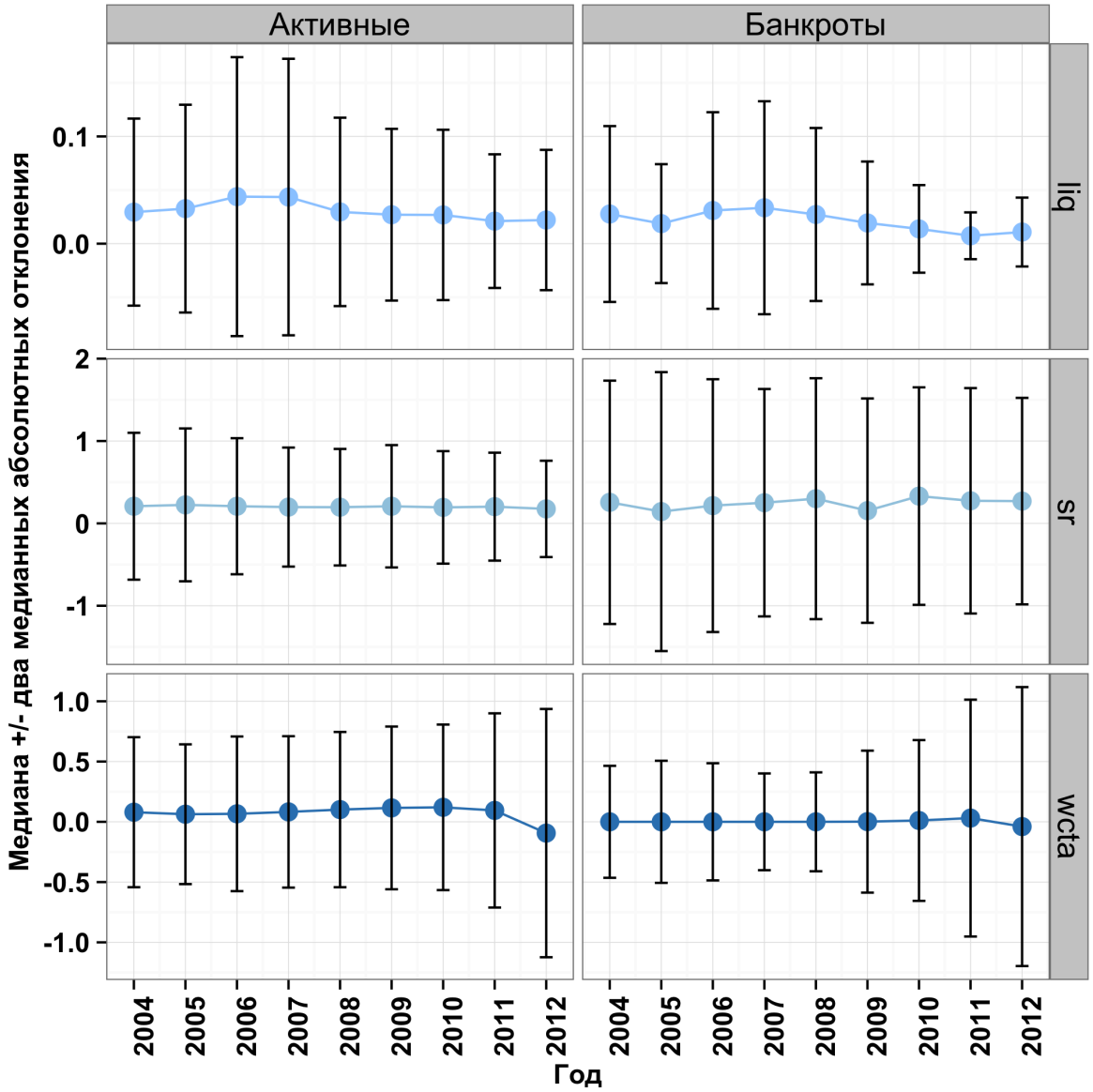


В группах приведено по три примера в иллюстративных целях, так как динамика остальных переменных в этих же группах качественно не отличается.

Подробное объяснение формул, по которым рассчитывались эти финансовые отношения, приведено в приложении, а также эти переменные будут вводиться по мере включения их в модели.

На графиках изображены медианы плюс/минус два медианных абсолютных отклонения. Для распределений, близких к нормальному, этот диапазон соответствовал бы 95%-ому доверительному интервалу.

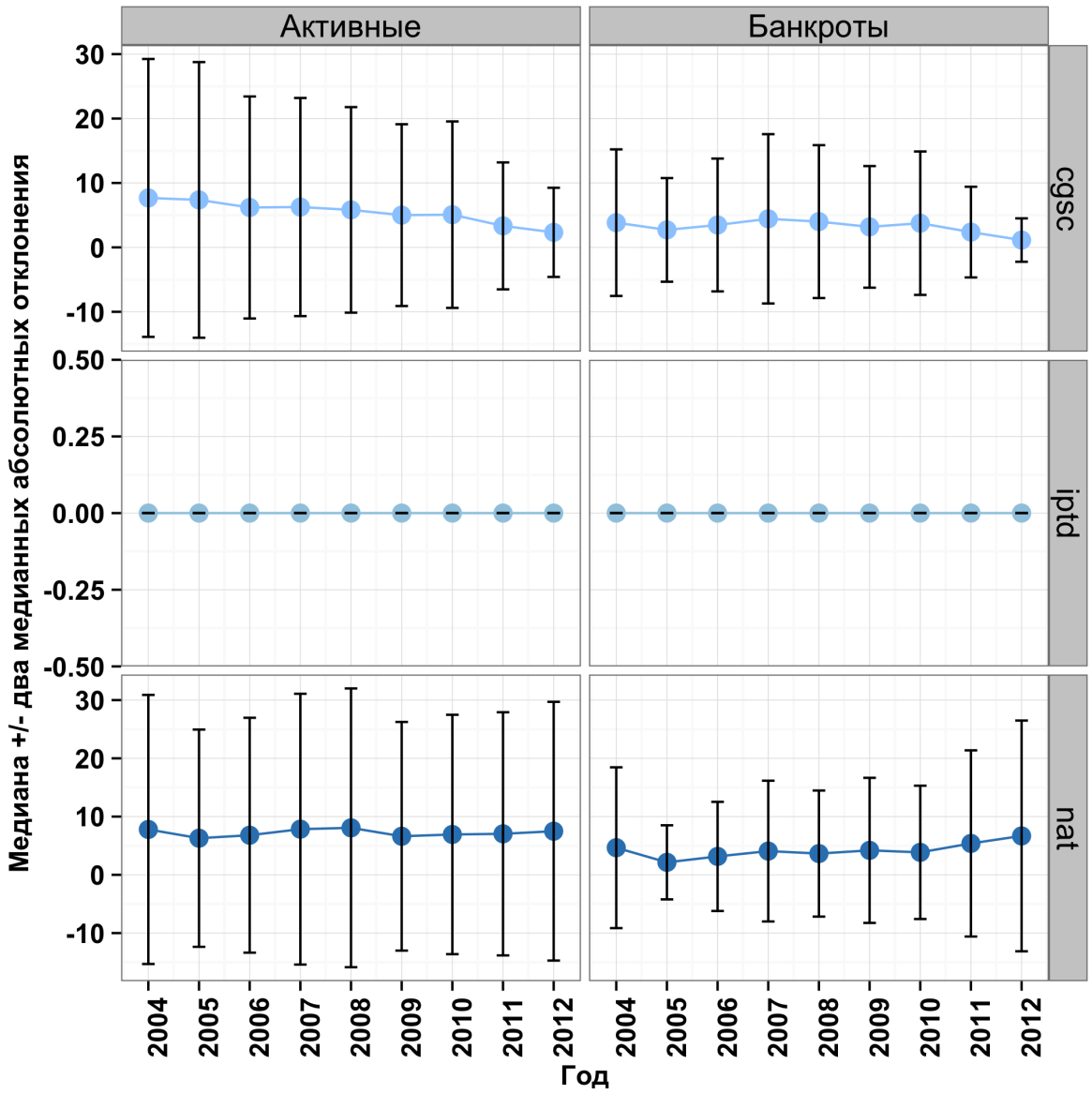
Рис. 12: Показатели ликвидности



Вне зависимости от группы переменных на графиках видно, что значения медианы любой взятой переменной для активных компаний и для

компаний-банкротов практически не отличается. Это является ещё одним индикатором в пользу нелинейной и сложной зависимости вероятности банкротства от объясняющих финансовых переменных.

Рис. 13: Показатели обслуживания долга



Отдельно стоит рассказать про переменные, где и медиана, и медианное абсолютное отклонение равны нулю. Эти переменные не константы, но так как более 50% наблюдений по этим переменным равны нулю, то медиана и медианное абсолютное отклонение соответственно тоже равняются нулю.

6 Теоретическое обоснование методов

Все методы данной работы можно условно разделить на четыре группы: логит- и пробит-модели, дискриминантный анализ, метод опорных векторов и классификационные деревья. Количество переменных в разных моделях разное, но для удобства все методы будут изложены для случая двух объясняющих переменных, x и z. Это позволяет избежать введения дополнительного индекса, обозначающего количество переменных, и сосредоточится на сути методов. Индекс года t также опущен для ясности изложения.

6.1 Логит- и пробит-модели

В логит- и пробит-моделях предполагается, что существует ненаблюдаемая переменная, склонность к банкротству, def_i^*, которая линейно зависит от регрессоров:

$$def_i^* = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i \tag{2}$$

Компания оказывается банкротом, если $def_i^* > 0$. Связь между наблюдаемым индикатором банкротства и ненаблюдаемой склонностью к банкротству имеет вид:

$$def_i = \begin{cases} 1, \text{ если } def_i^* > 0 \\ 0, \text{ если } def_i^* \leq 0 \end{cases} \tag{3}$$

Различие в логит- и пробит-моделях состоит в предположениях о распределении ε_i. В обоих случаях случайные ошибки предполагаются независимыми. В логит-модели предполагается, что ε_i имеет логистическое

распределение, а в пробит-модели — стандартное нормальное $N(0,1)$ распределение.

Оба типа моделей оцениваются с помощью метода максимального правдоподобия. Функция правдоподобия имеет вид:

$$L = \prod_i f(def_i|x_i, z_i) = \prod_{def_i=1} F(\beta_1 + \beta_2 x_i + \beta_3 z_i) \prod_{def_i=0} (1 - F(\beta_1 + \beta_2 x_i + \beta_3 z_i)), \quad (4)$$

где $F()$ — функция распределения ε_i, разная для логит- и пробит-моделей.

Оценки коэффициентов в логит- и пробит-моделях напрямую несравнимы. Логистическое распределение похоже на нормальное $N(0, \pi^2/3)$, поэтому оценки коэффициентов в логит-модели примерно равны оценкам коэффициентов в пробит-модели, домноженным на $\pi/\sqrt{3} \approx 1.8$.

6.2 Три вида дискриминантного анализа

В дискриминантном анализе предполагается, что вероятность банкротства отдельного предприятия равна $\mathbb{P}(def_i = 1) = p$ и специфицируется закон распределения регрессоров x_i и z_i при фиксированном y_i.

- В **линейном дискриминантном анализе** предполагается, что условное распределение регрессоров при фиксированном def_i — многомерное нормальное. Вектор средних может отличаться для банкротов и активных предприятий, а ковариационная матрица — одинаковая:

$$f(x_i, z_i|def_i = 0) \sim N(\mu_0, \Sigma) \quad (5)$$

$$f(x_i, z_i|def_i = 1) \sim N(\mu_1, \Sigma) \quad (6)$$

- В **квадратичном дискриминантном анализе** предполагается, что условное распределение регрессоров при фиксированном def_i — многомерное нормальное. И вектор средних, и ковариационная матрица могут отличаться для банкротов и активных предприятий:

$$f(x_i, z_i|def_i = 0) \sim N(\mu_0, \Sigma_0) \tag{7}$$

$$f(x_i, z_i|def_i = 1) \sim N(\mu_1, \Sigma_1) \tag{8}$$

- В **дискриминантном анализе смеси нормальных распределений** или смешанном дискриминантном анализе предполагается, что условное распределение регрессоров при фиксированном def_i — смесь трёх многомерных нормальных распределений с общей ковариационной матрицей:

$$f(x_i, z_i|def_i = 0) \sim p_{0a}N(\mu_{0a}, \Sigma) + p_{0b}N(\mu_{0b}, \Sigma) + p_{0c}N(\mu_{0c}, \Sigma) \tag{9}$$

$$f(x_i, z_i|def_i = 1) \sim p_{1a}N(\mu_{1a}, \Sigma) + p_{1b}N(\mu_{1b}, \Sigma) + p_{1c}N(\mu_{1c}, \Sigma) \tag{10}$$

Модели квадратичного дискриминантного анализа и анализа смеси распределений являются обобщениями модели линейного дискриминантного анализа.

Оцениваются все три модели с помощью метода максимального правдоподобия. Функция правдоподобия имеет вид:

$$L = \prod f(x_i, z_i, def_i) = \prod f(x_i, z_i|def_i)f(def_i) = p^{\sum def_i}(1-p)^{n-\sum def_i} \prod f(x_i, z_i|def_i), \tag{11}$$

где n — общее количество наблюдений в выборке.

6.3 Метод опорных векторов

В отличие от предыдущих методов, в методе опорных векторов не специфицируется вероятностная модель для объясняемой переменной или регрессоров.

Допустим в пространстве проведена гиперплоскость. Наблюдения, попавшие по одну её сторону, прогнозируются как банкроты, попавшие по другую сторону — как активные предприятия. Данная гиперплоскость называется разделяющей. Если дополнительно провести гиперплоскости на расстоянии $h/2$ от исходной, то получится разделяющая полоса шириной h.

Метод опорных векторов находит разделяющую полосу максимальной ширины с учётом штрафа за неправильно классифицированные наблюдения. Строго говоря, разделяющая гиперплоскость и ширина полосы h подбираются так, чтобы минимизировать функцию:

$$Q = \frac{2}{h^2} + \frac{C}{h} \sum_i d_i, \tag{12}$$

где d_i — расстояния от разделяющей плоскости до неправильно классифицированных наблюдений, C — параметр штрафа.

Гиперплоскость строится не в исходном пространстве регрессоров, а в преобразованном, так называемом **спрямляющем пространстве**. Для того, чтобы считать расстояния в спрямляющем пространстве достаточно знать формулу, задающую скалярное произведение в нём. В данной работе используется бесконечномерное гауссово спрямляющее пространство. Скалярное произведение двух наблюдений $r = (x, z)$ и $r' = (x', z')$ задаётся формулой:

$$k(r, r') = \exp(-\sigma |r - r'|^2) \tag{13}$$

Прогнозные вероятности получаются путём построения дополнительной логит-регрессии индикатора банкротства на расстояние (с учётом знака) между наблюдением и разделяющей гиперплоскостью в спрямляющем пространстве.

Подбор параметров C и σ осуществляется с помощью 10-кратной **кросс-валидации**:

1. Задаётся сетка значений параметров C и σ.
2. Для каждой пары значений C и σ:
   (a) Выборка делится случайным образом на 10 равных частей.
   (b) Поочерёдно выкидывается каждая из этих частей.
   (c) Модель оценивается по 9-ти оставшимся частям выборки, а прогнозы вероятностей строятся на выкинутую часть.
   (d) Таким образом, получается прогноз вероятности банкротства для каждого наблюдения.
3. Выбирается то значение параметров C и σ, при котором точность прогнозов максимальна.
4. Модель переоценивается с данными C и σ по всем имеющимся наблюдениям.

6.4 Классификационные деревья и случайный лес

Так же как и в методе опорных векторов, здесь не специфицируется вероятностная модель для объясняемой переменной или регрессоров.

Для каждого дерева определён **индекс Джини**, измеряющий «неидеальность» дерева. Индекс Джини — это вероятность того, что типы двух

предприятий (банкрот или активное) не совпадут, если первое выбирается случайно из всей выборки, а второе выбирается случайно из того же терминального узла дерева, что и первое. Формулой индекс Джини определяется как:

$$I_G(Tree) = \sum_i \frac{n_i}{n} 2h_i(1 - h_i), \tag{14}$$

где n — общий размер выборки, n_i — количество предприятий в i-ом терминальном узле, h_i — доля предприятий-банкротов в i-ом терминальном узле.

Для тривиального дерева из одного узла индекс Джини равен $I_G = 2h(1 - h)$, где h — доля предприятий-банкротов.

Для начала изложим **типичный алгоритм** построения отдельного классификационного дерева:

1. Посчитать стартовое значение индекса Джини как $I_G = 2h(1 - h)$.
2. Осуществить ветвление одного из терминальных узлов дерева на две части так, чтобы падение индекса Джини было максимальным.
3. Посчитать новое текущее значение индекса Джини.
4. Если терминальных узлов слишком много или в каждом терминальном узле слишком мало предприятий, то завершить процесс построения дерева. Иначе перейти к шагу 2.

Понятия «слишком много узлов» и «слишком мало предприятий в каждом узле» зависят от конкретной версии алгоритма. В использованной в данной работе версии алгоритма минимальное число предприятий в терминальном узле равно 5, а высота дерева (максимальная длина ветви) не больше 31.

Вероятность банкротства прогнозируется как доля банкротов в том терминальном узле, где лежит данное предприятие.

Алгоритм **случайного леса** состоит в построении большого количества, 500 в данной работе, классификационных деревьев для каждой модели. Особенности построения деревьев следующие:

1. Каждое дерево строится по случайной выборке с повторениями из исходной. Размер случайной выборки равен размеру исходной, но из-за возможности повторений некоторые элементы исходной выборки могут не использоваться при построении дерева.

2. Дерево строится до тех пор, пока возможно дальнейшее уменьшение индекса Джини. Фактически это означает, что в каждом терминальном узле остаются либо только банкроты, либо только активные предприятия.

3. При каждом ветвлении дерева случайным образом отбирается $\lfloor\sqrt{k}\rfloor$ штук из исходных k регрессоров. Затем из предварительно отобранных $\lfloor\sqrt{k}\rfloor$ регрессоров выбирается тот, который обеспечивает максимальное падение индекса Джини.

Вероятность банкротства прогнозируется как доля количества деревьев, «голосующих» за то, что данное предприятие будет банкротом.

6.5 Описание использованных пакетов

База данных «Руслана» не позволяет скачивать за один раз большой объём данных, поэтому для автоматизации сбора данных из базы использовался язык программирования `python` и библиотека `selenium` для автоматического управления браузером.

Дальнейшая обработка данных происходила в языке программирования `R` версии 3.0.2 с использованием графической оболочки `Rstudio`. Стандартный стиль работы в `R` предполагает, что весь массив данных сразу загружается в оперативную память компьютера. Однако в данной работе объём полученных данных оказался настолько велик, что работа с полным массивом данных в оперативной памяти на имеющемся компьютере оказалась невозможной.

Поэтому была организована SQL-база данных `SQLite` с рабочим объемом более 20 Гб. Поочерёдно таблички для каждого отдельного года загружались в `R` и анализировались далее. Для взаимодействия между средой `R` и базой `SQLite` использовался пакет `RSQLite`.

В силу многообразия применяемых методов использовался ряд дополнительных библиотек. Логит- и пробит-модели оценивались стандартными средствами `R`. Для оценивания случайного леса в `R` существуют два пакета, `party` и `randomForest`. Хотя пакет `party` более свежий, чем `randomForest`, скорость построение леса у него существенно ниже, поэтому в данной работе использовался пакет `randomForest`. Для построения классификационного дерева применялись пакеты `tree` и `rpart`. Оценивание с помощью метода опорных векторов реализовано в пакете `kernlab`. Он обладает скоростью работы, сравнимой со скоростью своего конкурента, пакета `e1071`. Подбор параметра штрафа C в методе опорных векторов осуществлялся с помощью кросс-валидации, реализованной в пакете `caret`. Для линейного и квадратичного дискриминантного анализа использовался пакет `MASS`, а смешанный дискриминантный анализ был проведён с помощью пакета `mda`.

7 Модели

В силу того, что данные годовые, строились отдельные модели для каждого года для прогнозирования дефолта на год вперёд, то есть брались текущие значения объясняющих переменных и будущее на следующий год значение зависимой переменной:

$$def_{it+1} = \mathbf{x}_{it}\beta_t + \varepsilon_{it}, \tag{15}$$

где i — идентификационный номер компании, t — период, def_{it+1} — это индикатор дефолта (равен 1, если компания i стала банкротом в году $t+1$, и равен 0 иначе), $\mathbf{x}_{it}$ — вектор-строка характеристик компании i в году t, β_t — вектор-столбец неизвестных параметров в году t. Далее для простоты индекс t будет опущен.

Более того, модели строились по каждой отрасли в отдельности и по двум типам выборок внутри одной отрасли: балансированной и небалансированнной.

В балансированных выборках содержалось равное число банкротов и активных компаний, причём были взяты все банкроты за период (как было показано в графическом анализе их гораздо меньше, чем активных компаний), и к ним добавлялось такое же количество случайным образом выбранных активных компаниях.

В небалансированных выборках опять же присутствовали все дефолтные компании, но случайно выбранных активных компаний было больше, а именно 20% от всех активных в тот период компаний.

Исследователи расходятся во мнениях по поводу влияния наличия неравного числа исходов в выборках, именно по этой причине модели построены и по балансированным, и по небалансированным выборкам.

20% всех активных предприятий взято по причине очень большого массива данных, и даже на этих массивах модели оценивались довольно долго. Количество наблюдений разнится от 110 до 56 210 наблюдений.

Стоит подчеркнуть, что такое небольшое число наблюдений в выборках как 200 связано с тем, что в данных оказалось очень много пропущенных значений, и прежде чем отбирать и балансировать/не балансировать массив данные были очищены от пропущенных значений по переменным, которые включались в модель.

7.1 Выбор переменных

Большое количество финансовых отношений затрудняло выбор переменных для построения моделей. В основном авторы либо опирались на выбор переменных в других исследованиях, либо применяли метод главных компонент, однако в данной работе были применены три варианта отбора переменных:

1. Выбор финансовых переменных, аналогичных модели Альтмана и Сабато для предприятий малого и среднего бизнеса, и добавление нефинансовых переменных;

2. Выбор финансовых переменных с помощью критерия частоты использования в других работах и добавление нефинансовых переменных;

3. Выбор переменных обоих типов с помощью LASSO.

Рассмотрим каждый из вариантов по отдельности.

7.1.1 Адаптация модели Альтмана и Сабато

В работе 2007 года Альтман и Сабато предложили модель, предназначенную именно для прогнозирования банкротств небольших фирм. Однако в силу огромного числа пропущенных значений в данных или вовсе отсутствия некоторых показателей были подобраны наиболее близкие к указанным в исследовании этих авторов переменные. Они представлены в таблице ниже.

Таблица 6: Финансовые переменные в модели Альтмана и в текущей модели

Модель Альтмана, Сабато	**Текущая модель**	**Причина замены (если была)**	**Название в данных**
$\frac{\text{EBITDA}}{\text{Total assets}}$	$\frac{\text{EBIT}}{\text{Total assets}}$	Пропуски в переменной амортизация	ebta
$\frac{\text{Short-term debt}}{\text{Book value of equity}}$	$\frac{\text{Short-term debt}}{\text{Total equity}}$	—	stdte
$\frac{\text{Retained earnings}}{\text{Total assets}}$	$\frac{\text{Net income}}{\text{Total assets}}$	Отсутствие переменной нераспределённая прибыль	roa
$\frac{\text{Cash}}{\text{Total assets}}$	$\frac{\text{Cash and cash equivalent}}{\text{Total assets}}$	—	liq
$\frac{\text{EBITDA}}{\text{Interest expenses}}$	$\frac{\text{Interest paid}}{\text{Total debts}}$	Пропуски в переменной амортизация	iptd
—	$\frac{\text{Total equity}}{\text{Total assets}}$	—	sr

Последняя в таблице переменная не присутствовала в модели Альтамана и Сабато, но была добавлена в текущую модель, так как этот показатель характеризует степень стабильности предприятия.

Однако в своей модели Альтман и Сабато использовали только нефинансовые переменные, так как в их базе данных не было качественных характеристик данных. В данной же работе наличие нефинансовых переменных рассматривается как важный фактор улучшения модели, поэтому во все модели включён как минимум один нефинансовый показатель, а именно возраст. С данными финансовыми отношениями есть модели двух типов.

Первый тип моделей содержит выбранные финансовые переменные и нефинансовую переменную, отвечающую за возраст компании.

$$def_i = \beta_1 + \beta_2 \cdot iptd_i + \beta_3 \cdot ebta_i + \beta_4 \cdot stdte_i + \beta_5 \cdot roa_i + \beta_6 \cdot liq_i \quad (16)$$
$$+ \beta_7 \cdot sr_i + \beta_8 \cdot age_i + \varepsilon_i,$$

где def_i — это def_{it+1}, то есть дефолт в периоде $t+1$, где опущен индекс $t+1$, а у всех остальных переменных опущен индекс t.

Такая постановка модели была сделана из-за того, что некоторые методы, а именно линейный и квадратичный дискриминантный анализ и дискриминантный анализ смеси распределений требуют использование только числовых переменных. Из всех нефинансовых числовой переменной является только возраст фирмы, поэтому он добавлен в эту модель. Таким образом, этот тип моделей оценён с помощью всех применяемых в данной работе методов: трёх видов дискриминантного анализа, логит- и пробит-моделей, метода опорных векторов, метода классификационных деревьев и алгоритма случайного леса.

Второй тип моделей содержит финансовые и нефинансовые переменные, среди которых не только возраст, но и федеральный округ, последний доступный размер и организационная форма.

$$def_i = \beta_1 + \beta_2 \cdot iptd_i + \beta_3 \cdot ebta_i + \beta_4 \cdot stdte_i + \beta_5 \cdot roa_i + \beta_6 \cdot liq_i \quad (17)$$
$$+ \beta_7 \cdot sr_i + \beta_8 \cdot age_i + \beta_9 \cdot fedreg_i + \beta_{10} \cdot lasize_i + \beta_{11} \cdot legform_i + \varepsilon_i$$

Здесь для компактности записи для нефинансовых качественных переменых используется выражение вида $\beta_9 \cdot fedreg_i$, что на самом деле означает включение соответствующего количества дамми-переменных.

Из-за особенностей дискриминантного анализа этот тип моделей оценён с помощью всех остальных методов кроме него.

7.1.2 Частота использования в исследованиях

Второй способ, который применялся для отбора финансовых показателей, — частота упоминаний в исследованиях дефолтов компаний в разных странах. Для этого была взята статья Белловари, Жиакомино и Акерса (2007), в которой приведено сравнение статей по прогнозированию банкротств компаний с 1930 по 2007 год.

Критерием для выбора стало использование показателя в более, чем в 27 исследованиях. Стоит сказать, что максимальное число упоминаний 54 раза, и 27 было выбрано таким образом, чтобы в моделях встречались только наиболее широко распространённые показатели.

В таблице 7 представлены отношения, выбранные для анализа, и сколько раз они использовались при прогнозировании банкротств.

Таблица 7: Финансовые переменные, упорядоченные по популярности

Финансовое отношение	**Частота упоминаний (раз)**	**Название в данных**
$\frac{\text{Net income}}{\text{Total assets}}$	54	roa
$\frac{\text{Current assets}}{\text{Current liabilities}}$	51	cr
$\frac{\text{Working capital}}{\text{Total assets}}$	45	wcta
$\frac{\text{EBIT}}{\text{Total assets}}$	35	ebta
$\frac{\text{Sales}}{\text{Total assets}}$ [a]	32	—
$\frac{\text{Current assets - stocks}}{\text{Current liabilities}}$	30	lr
$\frac{\text{Total debt}}{\text{Total assets}}$	27	tdta
$\frac{\text{Interest paid}}{\text{Total debt}}$	—	iptd

[a] Из-за большого числа пропусков в переменной продажи это отношение заменено на Interest paid/Total debt как показатель, характеризующий долг

Аналогично прошлому способу выбора переменных при данном наборе финансовых показаталей строились два типа моделей. Первый тип с

финансовыми отношениями и возрастом компаний, который оценивался всеми методами, используемыми в данной работе.

$$def_i = \beta_1 + \beta_2 \cdot iptd_i + \beta_3 \cdot roa_i + \beta_4 \cdot cr_i + \beta_5 \cdot wcta_i + \beta_6 \cdot ebta_i \quad (18)$$
$$+ \beta_7 \cdot lr_i + \beta_8 \cdot tdta_i + \beta_9 \cdot age_i + \varepsilon_i$$

Второй тип моделей — модели с добавлением других нефинансовых переменных кроме возраста — оценивался методами кроме видов дискриминантного анализа.

$$def_i = \beta_1 + \beta_2 \cdot iptd_i + \beta_3 \cdot roa_i + \beta_4 \cdot cr_i + \beta_5 \cdot wcta_i + \beta_6 \cdot ebta_i \quad (19)$$
$$+ \beta_7 \cdot lr_i + \beta_8 \cdot tdta_i + \beta_9 \cdot age_i + \beta_{10} \cdot fedreg_i + \beta_{11} \cdot lasize_i$$
$$+ \beta_{12} \cdot legform_i + \varepsilon_i$$

7.1.3 LASSO

Третий способ выбора переменных — применение LASSO (Least Absolute Shrinkage and Selection Operator). Однако в отличие от предыдущих вариантов, где выбор производился только среди финансовых переменных, здесь отбор производился из всех имеющихся в массиве данных переменных.

LASSO в формальном виде можно записать как задачу минимизации

$$\min_{\beta}(y - X\beta)^T(y - X\beta) \quad (20)$$

при условии

$$\sum_{j=1}^{p} |\beta_j| \leq t \quad (21)$$

Используя лагранжиан, можно записать задачу таким образом:

$$\min_{\beta}(y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (22)$$

Ограничение такого вида очень полезно для данной работы, потому что оно само помогает выбрать спецификацию модели, определяя наиболее важные переменные. Минимизировались лагранжианы для различных лет, причём для каждого года брался весь имеющийся массив данных, а пропущенные переменные были заполнены средними значениями. Результат, который получился, оказался довольно стабильным. Около 80% отобранных переменных совпадали вне зависимости от года. В таблице 8 представлены только финансовые переменные, отобранные этим способом.

Таблица 8: Финансовые переменные, отобранные LASSO

Финансовое отношение	**Название в данных**
$\frac{\text{Non-current liabilities + Loans}}{\text{Total equity}}$	gg
$\frac{\text{Turnover}}{\text{Total equity + Non-current liabilities}}$	nat
$\frac{\text{EBIT}}{\text{Turnover}}$	ebtm
$\frac{\text{Long-term debt}}{\text{Total assets}}$	ltdta
$\frac{\text{Working capital}}{\text{Total assets}}$	wcta

Однако стоит отметить, что LASSO выбрал все нефинансовые исследуемые переменные как важные, что ещё раз подтверждает правомерность их использования в моделях. Как и в двух предыдущих случаях были две модели. Первый тип с финансовыми переменными и возрастом, который оценивался всеми методами.

$$def_i = \beta_1 + \beta_2 \cdot gg_i + \beta_3 \cdot nat_i + \beta_4 \cdot ebtm_i + \beta_5 \cdot ltdta_i + \beta_6 \cdot wcta_i + \beta_9 \cdot age_i + \varepsilon_i \quad (23)$$

И второй тип моделей — модель со всеми выбранными финансовыми и нефинансовыми показателями. Оценивание производилось, как и при

предыдущих наборах, без дискриминантного анализа.

$$def_i = \beta_1 + \beta_2 \cdot gg_i + \beta_3 \cdot nat_i + \beta_4 \cdot ebtm_i + \beta_5 \cdot ltdta_i + \beta_6 \cdot wcta_i \quad (24)$$
$$+ \beta_7 \cdot age_i + \beta_8 \cdot fedreg_i + \beta_9 \cdot lasize_i + \beta_{10} \cdot legform_i + \varepsilon_i$$

7.2 Другие характеристики моделей

Итак, для каждого года с 2004 по 2012 и внутри года для каждой отрасли были построены модели по шести описанным выше формулам по балансированным и небалансированным выборкам. В таблице 9 представлено количество моделей в зависимости от типа выборки, формулы и метода оценивания.

Таблица 9: Сводная таблица моделей

Критерий выбора финансовых переменых	**Нефинансовые переменные**	**Количество моделей внутри метода**							
		ЛДА	**КДА**	**СДА**	**Логит**	**Пробит**	**Метод опорных векторов**	**Дерево**	**Лес**
Балансированные выборки									
Альтман и Сабато	Возраст	36	36	36	36	36	36	36	36
	Все				36	36	36	36	36
Популярность	Возраст	36	36	36	36	36	36	36	36
	Все				36	36	36	36	36
LASSO	Возраст	36	36	36	36	36	36	36	36
	Все				36	36	36	36	36
Количество моделей		108	108	108	216	216	216	216	216
Небалансированные выборки									
Альтман и Сабато	Возраст	36	36	36	36	36	31	36	36
	Все				36	36	31	36	36
Популярность	Возраст	36	36	36	36	36	30	36	36
	Все				36	36	30	36	36
LASSO	Возраст	36	36	36	36	36	30	36	36
	Все				36	36	30	36	36
Количество моделей		108	108	108	216	216	182	216	216
Итого количество моделей		216	216	216	432	432	398	432	432

Всего было оценено 2774 модели. Стоит отметить, что в небалансированных выборках большое количество наблюдений не позволило оценить некоторые модели с помощью метода опорных векторов и применения кросс-валидации (подробное объяснение будет приведено немного ниже). Поэтому их 30 или 31 вместо 36 (9 лет по 4 отрасли) в зависимости от

формулы, по которой оценивалась модель. Все эти неоценённые модели относились к отрасли оптовой и розничной торговли, так как именно там было самое большое число наблюдений по сравнению с другими отраслями все зависимости от года, что можно было видеть из графического анализа.

Ввиду большого количества как самих моделей, так и наблюдений в рамках отдельных моделей, оценивание моделей требовало большого количества времени. Для того, чтобы ускорить процесс оценивания был использован виртуальный компьютер в вычислительном облаке компании Амазон (Amazon Elastic Compute Cloud, `aws.amazon.com`).

Из множества доступных конфигураций виртуального компьютера была выбрана `c3.2xlarge`. Это виртуальный компьютер с 15 Гб оперативной памяти, оптимизированный для параллельных вычислений, мощностью в 8 условных базовых (8 ECU, EC2 Compute Unit).

На виртуальную машину был поставлен образ с операционной системой Ubuntu 13.10 и сервером R-studio 0.98.501, взятый с сайта `www.louisaslett.com/RStudio_AMI/`. Для одновременного функционирования нескольких ядер процессора использовался пакет `doMC`. По факту оптимальное быстродействие достигалось при использовании 4-х ядер.

На этом виртуальном компьютере оценивались алгоритмы, которые хорошо распараллеливаются, то есть алгоритмы, разные части которых могут одновременно реализовываться на разных ядрах процессора. Из рассматриваемых алгоритмов — это случайный лес и метод опорных векторов с кросс-валидацией. Каждый случайный лес включал 500 деревьев, поэтому каждое ядро процессора оценивало 125 деревьев. Для кросс-валидации метода опорных векторов был задействован пакет `caret`.

К сожалению, из-за ошибки в пакете `caret`, проявлявшейся иногда при использовании нескольких ядер, оценивание некоторых моделей производилось только на одном ядре процессора. Это существенно увеличило время работы и сделало недоступным на практике оценивание нескольких моделей методом опорных векторов с кросс-валидацией. Описание ошибки есть на страничке `stackoverflow.com/questions/11059925/`.

С учётом времени загрузки базы данных (больше 20 Гб), оценивания моделей и выгрузки полученных результатов суммарное время работы виртуального компьютера составило более 50 часов. Оплата за использование сервиса составила около 25 долларов.

Рис. 14: Наблюдения в моделях

Рис. 14а: Количество наблюдений в балансированных моделях

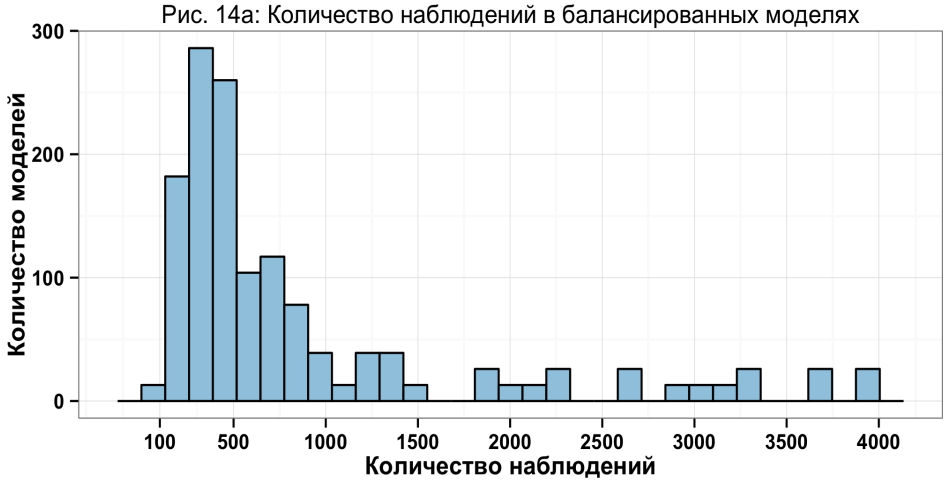
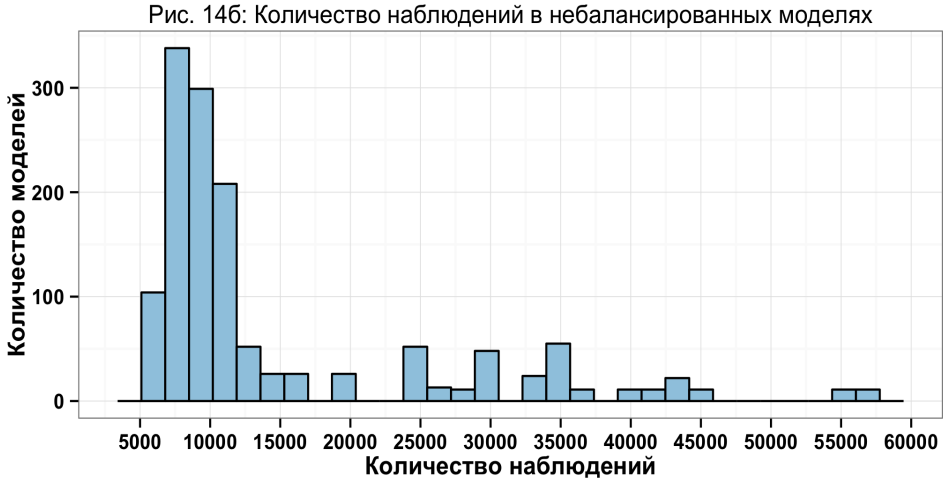


Рис. 14б: Количество наблюдений в небалансированных моделях



Также важно сказать, что из-за наличия пропущенных переменных для разных лет, отраслей, типа выборки и формулы количество наблюдений в разных моделях отличалось. На графиках 14а и 14б показаны количество наблюдений и соответствующее им число моделей.

В балансированных выборках количество наблюдений лежит в интервале от 100 до 4 000, причём большинство таких моделей имело в сумме около 500 наблюдений, то есть 250 дефолтных и 250 активных компаний. В небалансированных выборках насчитывалось от 5 000 до около 60 000 наблюдений, в основном, наблюдений было около 10 000.

8 Прогнозы

Ввиду того, что цель работы — прогнозирование вероятности наступления серьёзных финансовых трудностей, после оценивания моделей были построены прогнозы. Все прогнозы делались только на год вперёд, хотя некоторые авторы прогнозировали не только на год, но и на два года, и даже на пять вперёд.

Однако предполагать, что текущее состояние компании поможет предсказать, что будет с ней через пять, кажется не очень реалистичным, ведь положение может измениться не то, что в месяцы, а в считаные дни. Но практически у всех компаний отчётность подаётся не чаще, чем раз в квартал. Недоступность квартальной отчётности и убеждение в том, что текущий год значительно влияет именно на наиближайший следующий год, выступают в пользу построения прогнозов лишь на год вперёд.

Итак, например, если модель была построена по данным 2004 года (то есть данные по объясняющим переменным датировались 2004 годом, а зависимая переменная — 2005 годом), то прогноз делался на 2005 год (то есть то есть данные по объясняющим переменным датировались 2005 годом, а зависимая переменная — 2006 годом). На 2005 год брался весь массив данных, очищался от пропущенных значений по переменным, входящим в модель, а потом на очищенном массиве строился прогноз.

Прогнозы были построены по шести формулам для каждой из четырёх отраслей с помощью восьми методов и по двум типам выборок. Таким образом, были получены прогнозы на период с 2005 по 2012 год. Количество наблюдений в тестовых выборках составило от 26 590 до 273

600 наблюдений.

По причине того, что несколько моделей не были оценены из-за большого числа наблюдений, прогнозы по ним соответственно отсутствуют. По моделям 2012 года не было построено прогнозов из-за отсутствия данных за 2013 год. Однако для интерпретации коэффициентов в моделях и для более подробного представления построенных моделей будет использован именно 2012 год.

8.1 Критерии сравнения прогнозов

Для удобства введём стандартную терминологию, принятую в литературе по оценке качества моделей бинарной классификации. Исходы делятся на **положительные** и **отрицательные**. Положительные исходы в данной работе означают предприятия-банкроты, а отрицательные — активные компании.

Для наглядности приведём табличку 10:

Таблица 10: Матрица сопряжённости

		Фактический исход	
		Положительный	**Отрицательный**
Прогнозируемый исход	**Положительный**	Верно положительно True positive (TP)	Ложно положительно False positive (FP)
	Отрицательный	Ложно отрицательно False negative (FN)	Верно отрицательно True negative (TN)

Рассмотрим три основных показателя качества модели:

- **Точность** или доля верных классификаций (Accuracy, ACC) — отношение количества верных классификаций к общему количеству наблюдений.

- **Чувствительность** или доля верных положительных классификаций (True positive rate, TPR) — отношение количества верных по-

ложительных классификаций к общему количеству положительных исходов (банкротов).

- **Специфичность** или доля верных отрицательных классификаций (True negative rate, TNR) — отношение количества верных отрицательных классификаций к общему количеству отрицательных исходов (активных компаний).

Приведём математическеи формулировки:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{25}$$

$$TPR = \frac{TP}{TP + FN} \tag{26}$$

$$TNR = \frac{TN}{TN + FP} \tag{27}$$

Все рассматриваемые модели позволяют не просто получать прогноз вида банкрот-небанкрот, но и оценивать вероятность банкротства. Следовательно, конкретный прогноз зависит от порога для вероятности, за которым предприятие классифицируется как банкрот. Поэтому от выбора порога зависит и чувствительность, и специфичность, и точность. Чтобы изобразить зависимость чувствительности и специфичности от порога используют ROC-кривую (Receiver operating characteristic curve).

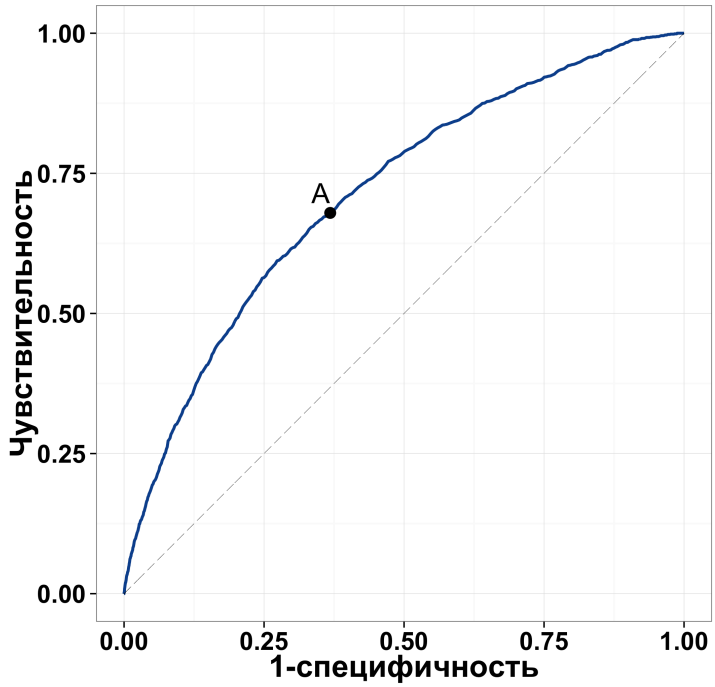
Для примера рассмотрим ROC-кривую для прогноза на 2012 год для модели по оптовой и розничной торговле, балансированным данным 2011 года и формуле, где финансовые переменные выбраны по критерию популярности их использования, а нефинансовые переменные представлены всеми имеющимися в данных.

Каждая точка на кривой ROC задаёт возможные значения чувстви-

тельности и специфичности для некоторого порога. По вертикали откладывается чувствительность, т.е. вероятность правильно классифицировать предприятие-банкрот, а по горизонтали — единица минус специфичность, то есть единица минус вероятность правильно классифицировать активное предприятие.

Например, точка A означает на графике 15, что при некотором пороге чувствительность будет равна 0.679, а специфичность — 0.632. Сами пороги на графике ROC не видны.

Рис. 15: Пример ROC-кривой. Алгоритм случайного леса



Увеличение порога означает движение влево вниз по кривой ROC. При пороге, равном нулю, все предприятия признаются банкротами (верхний правый угол графика) и чувствительность равна единице, а специфичность — нулю. При пороге, равном единице, все предприятия признаются активными (нижний левый угол графика) и чувствительность равна нулю, а специфичность — единице.

Если модель A лучше модели B при любом пороге, то при произвольной одинаковой чувствительности, в модели A выше специфичность, чем

в модели B. В этом случае ROC-кривая для модели A лежит левее и выше ROC-кривой для модели B. Например, на графике 16 видно превосходство алгоритма случайного леса над логит-моделью на данных по отрасли оптовой и розничной торговли в 2011 году.

Рис. 16: Сравение ROC-кривых

Рис. 16а: Случайный лес

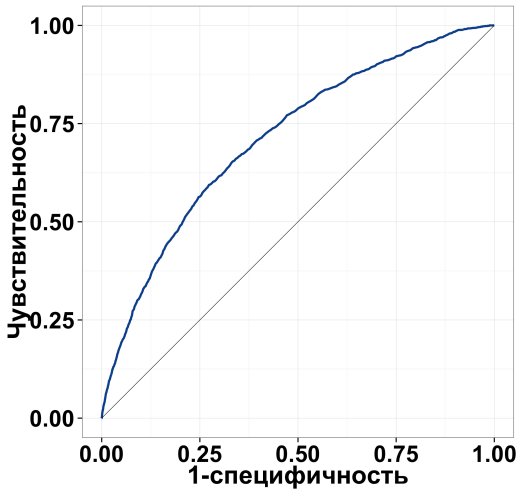
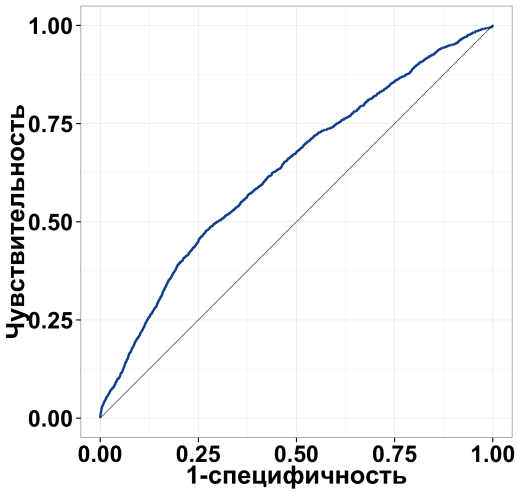


Рис. 16б: Логит



Поскольку количество предприятий-банкротов существенно меньше количества активных предприятий в любой отрасли в любом году, даже простейшая модель «признаём все предприятия активными» даёт очень высокую, близкую к единице, точность. К сожалению, такая модель лишена практической ценности для банка, решающего вопрос о кредитовании компании. Поэтому вопрос выбора наилучшей модели не может решаться на основании точности.

Вместо точности в этой работе используются два критерия:

- Площадь под ROC-кривой (Area Under Curve, AUC). Если модель A при любом пороге лучше модели B, то площадь по ROC-кривой в модели A будет больше площади в модели B. Также можно интепретировать величину AUC как вероятность того, что у случайно выбранной компании-банкрота спрогнозированная вероятность

банкротства будет выше, чем у случайно выбранной активной компании.

- Специфичность при заданном уровне чувствительности. Для банка, рассматривающего вопрос кредитования организации, выдача кредита фирме, которая обанкротится, является существенно более серьёзной ошибкой, чем невыдача кредита фирме, которая не обанкротится. То есть прежде всего банку интересен уровень чувствительности модели, то есть вероятность того, что будет верно классифицировано предприятие-банкрот. А затем уже, при некотором приемлемом уровне чувствительности, банк заинтересован в максимизации специфичности. Выбор приемлемого уровня чувствительности зависит от издержек и политики банка. В данной работе сравнивается специфичность моделей при чувствительности, равной 0.9. На кривой ROC данный критерий определяется точкой, находящейся на высоте 0.9.

Следует ещё раз отметить, что показатели качества можно считать тремя способами:

- Оценить модель по данным 2011 года, построить прогнозы на 2012 год.
- Оценить модель по данным 2011 года, построить прогнозы на 2011 год по той же выборке.
- Поделить выборку 2011 года на две части: обучающую и тестовую. Оценить модель по обучающей выборке, построить прогноз на 2011 год по тестовой выборке.

Показатели качества модели будут выше всего у второго способа, это некорректный способ, так как, включив в модель избыточно много переменных, можно достигнуть иллюзии высокого качества прогнозов. Третий способ является корректной версией второго. Мы выбираем первый способ, потому что именно с такой задачей сталкивается сама компания или кредитующая её организация.

8.2 Сравнение прогнозов

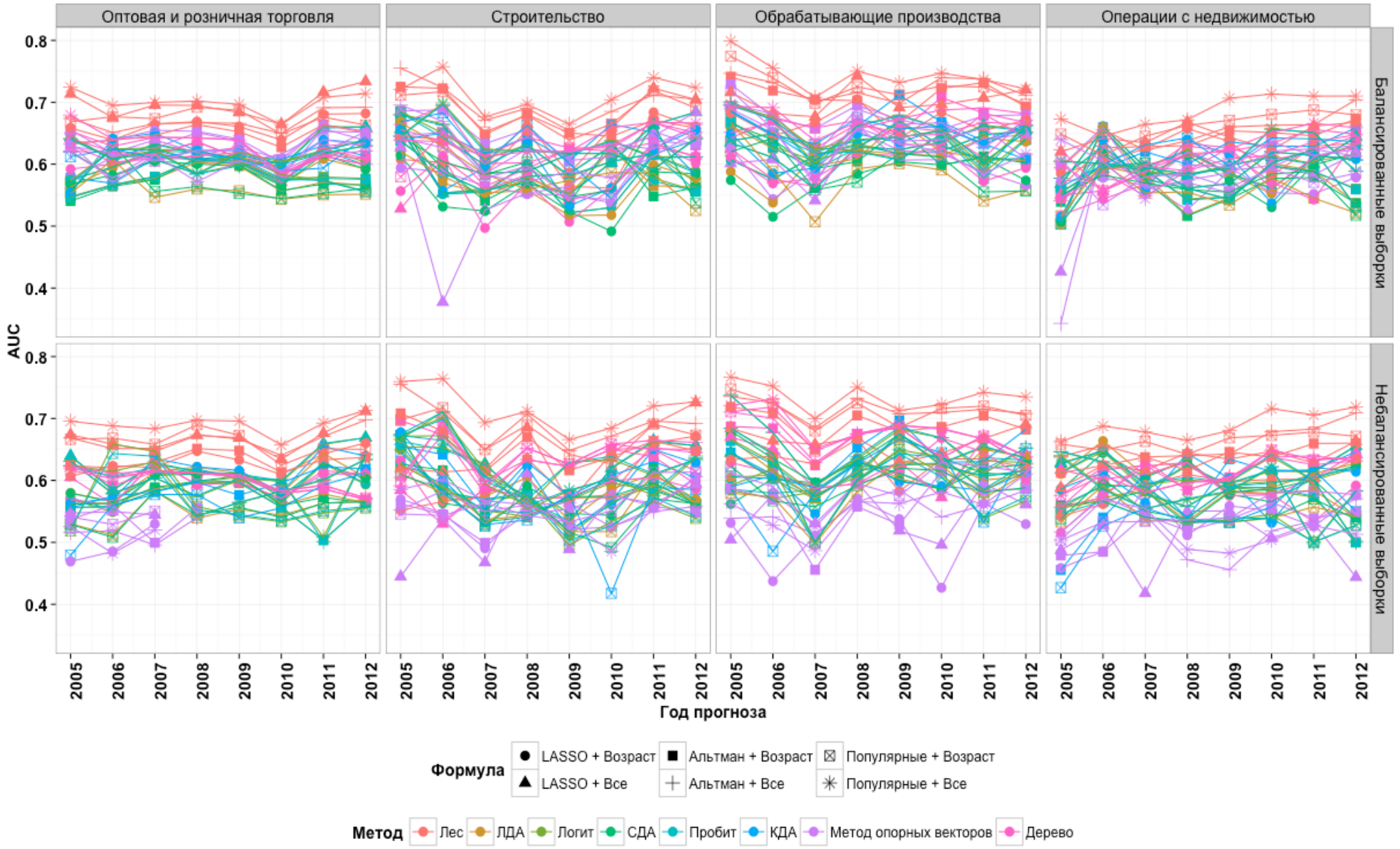
Посмотрим на результаты моделей по каждому из описанных выше критериев. Начнём с AUC. На следующей странице представлен график 17, отображающий значение AUC для моделей по всем шести формулам, по четырём отраслям, обоим типам выборок и с помощью применяемых в работе методов.

С первого взгляда кажется, что понять что-то по этому графику крайне тяжело. Однако с его помощью можно сделать несколько крайне важных для работы выводов.

Для прогнозов по моделям, построенным на балансированных выборках, практически всегда (за исключением трёх прогнозов на 2005 год с помощью метода опорных векторов) AUC превышал 0.5, что само по себе уже важный результат, так как это означает, что выбранные переменные оказывают влияние на вероятность дефолта в следующем году.

Что касается небалансированных выборок, стоит отметить, что в общем результаты схожи с AUC для моделей, построенных на балансированных выборках. Однако у метода опорных векторов есть значительные выбросы: AUC меньше 0.5. То есть можно сделать вывод, что наличие в выборке неравного числа исходов не оказывает значительного влияния

Рис. 17: Площадь под ROC-кривой для всех оценённых моделей



на то, насколько хорошо модель будет прогнозировать.

Более того, **алгоритм случайного леса** превосходит все остальные методы вне зависимости от отрасли, типа выборки и года. Как было отмечено в графическом анализе, нефинансовые переменные демонстрировали нелинейную зависимость с дефолтом, а алгоритм случайного леса как раз смог её отловить и учесть.

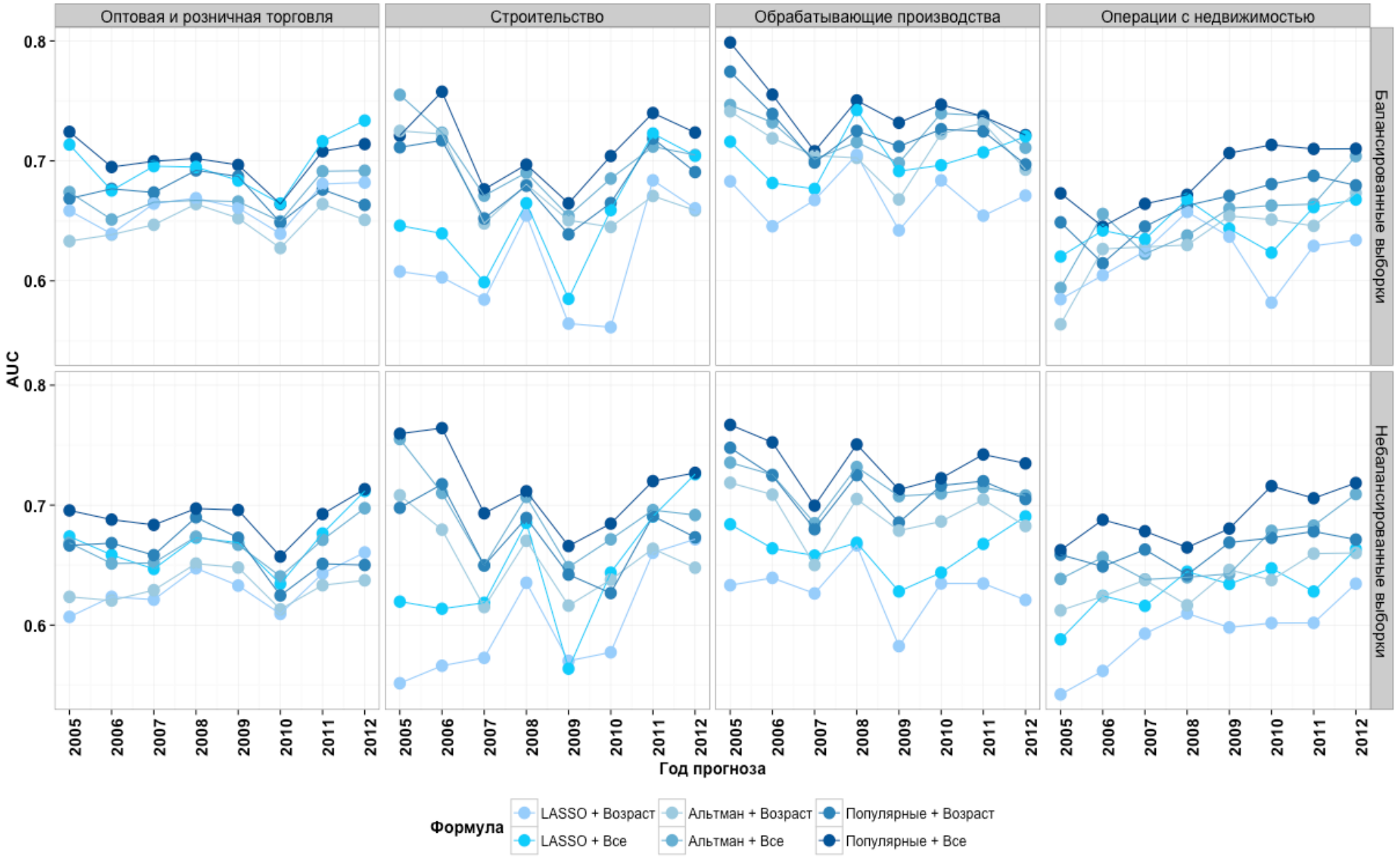
К тому же стоит отдельно остановиться на результатах леса для разных отраслей. В строительной отрасли и отрасли обрабатывающих производств AUC для случайного леса зачастую лежит в промежутке от 0.7 до 0.8. Но стоит отметить, что нельзя сказать, что для какой-то из исследуемых отраслей модели оказались значительно лучше, чем для других.

На втором месте по величине AUC часто следует классификационное дерево, которое тоже учитывает нелинейные зависимости. Логично, что оно хуже случайного леса, потому что в нём строится пятьсот деревьев, в то время как в классификационном дереве оно всего одно. Как и ожидалось, линейные методы прогнозирования банкротства предсказывают хуже.

Этот график также показывает **структурные сдвиги**. Особенно хорошо это можно видеть для строительной отрасли в 2007 и 2009 годах, когда все модели 2006 и 2008 годов вне зависимости от метода продемонстрировали падение AUC. То же самое можно наблюдать для обрабатывающей промышленности в 2007 году. Это означает, что зависимость между вероятностью дефолта и используемыми в моделях объясняющими переменными изменилась.

Однако по этому графику тяжело сказать о том, какой же подход к выбору переменных оказался наиболее успешным, хотя самая высокая

Рис. 18: Площадь под ROC-кривой для алгоритма случайного леса



кривая из AUC для случайного леса получена по модели, где были взяты финансовые переменные по их упоминаемости в работах, а также все нефинансовые переменные, которые были в данных.

Для рассмотрения метода, который показал наивысшие AUC, — случайного леса — более подробно и для дальнейшего обсуждения включённых переменных будет полезен график 18. Он показывает значения AUC только для моделей, построенных с помощью алгоритма случайного леса. Аналогичные графики для остальных методов приведены в приложении.

Здесь уже видно, что модель, куда входят все нефинансовые переменные и финансовые переменные, отобранные по частоте упоминаний в исследованиях прогнозирования банкротства, стабильно имеет большие значения площади под ROC-кривой. Однако для небалансированных выборок это первенство верно всегда, а вот для балансированных выборок есть несколько исключений.

Если сравнивать модели с одинаковыми финансовыми переменными, но различным набором нефинансовых характеристик, то модели, где учтены не только возраст компаний, но и федеральный округ, размер компании и организационная форма, всегда имеют лучшие прогнозы, чем модели, где присутствует только возраст. Это важный результат, потому что он доказывает, что нефинансовые переменные улучшают прогнозную силу моделей и их обязательно необходимо учитывать.

К тому же стоит отметить, что чёткого деления формул на первое, второе, третье места нельзя назвать, но для большинства моделей всё же на втором месте в основном идут модели с финансовыми переменными, аналогами переменным из модели Альтмана и Сабато. В то же

время модели с переменными, отобранными с помощью LASSO, иногда опережают, а иногда идут на последнем месте.

Относительно отраслей важно сказать, что все модели оказались более подходящими для обрабатывающей промышленности, потому что в балансированных выборках минимальный AUC равен 0.65 и максимальный AUC, равный 0.8, встречается только в моделях для этой отрасли. В то же время при небалансированных выборках разброс значений больше, и наибольшее значение ниже.

Теперь отдельно посмотрим на значения AUC, группированные по отраслям, методам и типам выборок, но не учитывая год. График 19, характеризующий это, приведён на следующей странице.

Чёрными горизонтальными линиями выделены медианы. Несомненное превосходство алгоритма случайного леса для всех отраслей здесь ещё более ярко выражено.

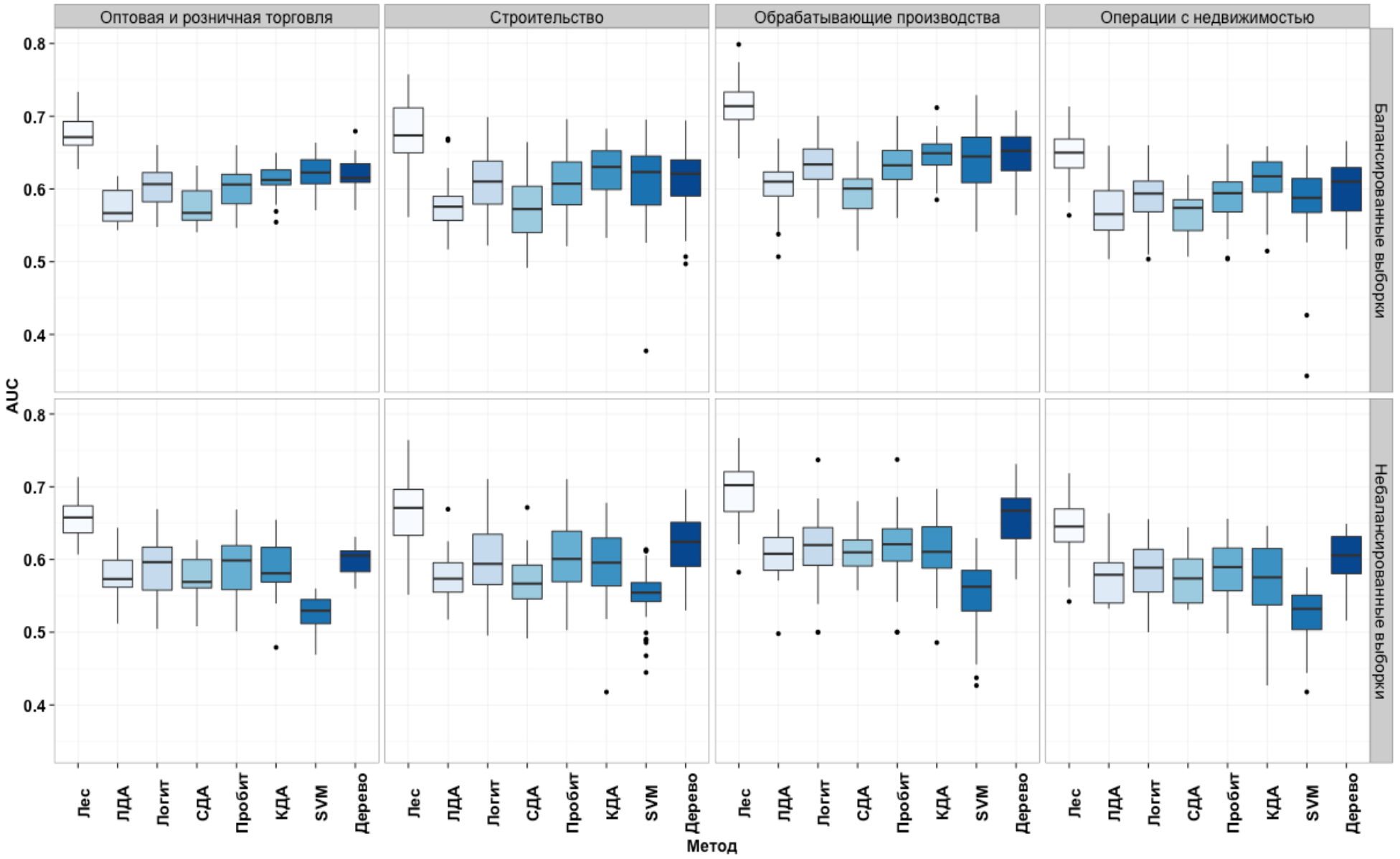
Для балансированных выборок наименьшая медиана наблюдается у линейного дискриминантного анализа, а на втором месте после леса следуют квадратичный дискриминантный анализ и дерево.

Для небалансированных выборок худшим оказался метод опорных векторов с применением кросс-валидации, а на втором месте после леса для всех отраслей оказалось дерево.

Однако различия в медианах для балансированных выборок и для небалансированных выборок не велики, все они выше 0.5. Для большинства моделей AUC колеблется в промежутке от 0.55 до 0.65.

Хотя этот график помогает глобально сравнить методы между собой, на нём нельзя выявить структурные сдвиги, что было возможно на предыдущих графиках, где по оси x были изображены года.

Рис. 19: Распределение площади под ROC-кривой в зависимости от метода



Также стоит отдельно более подробно рассмотреть влияние добавления в модель нефинансовых переменных на значение AUC. Ранее было отмечено, что для случайного леса, действительно, заметно превосходство моделей со всеми нефинансовыми переменными. Это верно и для других методов.

Чтобы это показатать, приведён график 20, показывающий значения AUC и медианы для каждого набора объясняющих переменных для каждой отрасли и типа выборки.

Если сравнивать по две модели, то есть при одном наборе финансовых объясняющих переменных сравнивать добавление возраста и добавление других нефинансовых характеристик, то можно заметить, что медианы AUC для моделей со всеми нефинансовыми показателями выше, чем при наличии одного возраста.

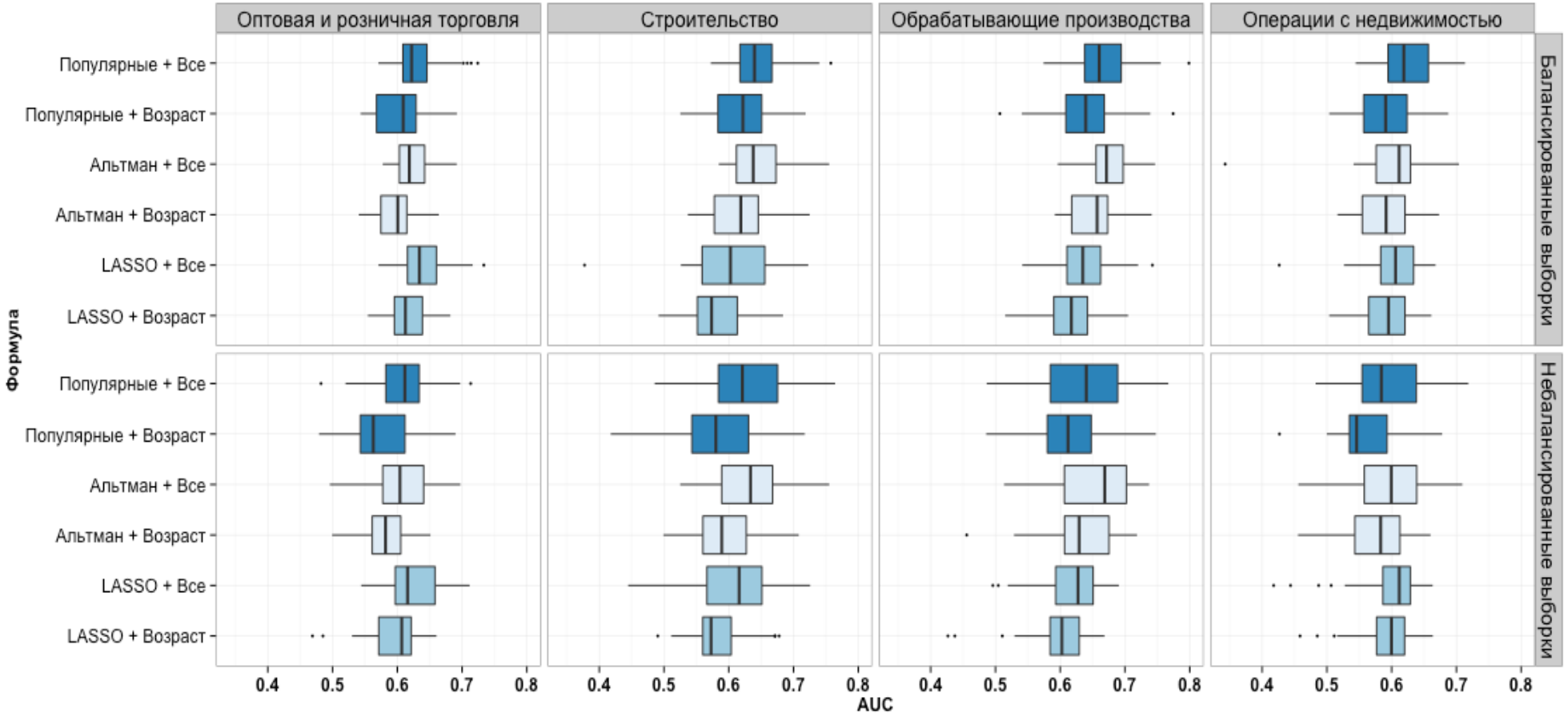
По сравнению с графиком 19 для деления по типу формулы характерна медиана, равная 0.6, причём это едино и для балансированных выборок, и для небалансированных выборок.

Отличие в медианах кажется не таким значительным, но наименьшее значение прощади под ROC-кривой для моделей со всеми нефинансовыми переменными всегда выше, чем для моделей с возрастом. То же самое характерно и для наибольшего AUC.

Более того, разброс значений AUC для моделей, построенных на небалансированным выборкам, больше, чем для моделей, построенных на балансированных выборках.

Однако недостатком данного графика является то, что здесь не видно явного превосходства одного набора финансовых переменных над другим.

Рис. 20: Распределение площади под ROC-кривой в зависимости от формулы



Рассмотрев в отдельности деление AUC по методам и по формулам, также приведём таблицу по среднему значению AUC внутри каждого метода и формулы.

Таблица 11: Средние AUC по методам и моделям

Критерий выбора финансовых переменых	**Нефинансовые переменные**	**Среднее значение AUC в зависимости от метода и формулы**							
		ЛДА	**КДА**	**СДА**	**Логит**	**Пробит**	**Метод опорных векторов**	**Дерево**	**Лес**
Балансированные выборки									
Альтман и Сабато	Возраст	0.5855	0.6275	0.5840	0.5929	0.5917	0.6363	0.6234	0.6666
	Все				0.6200	0.6192	0.6232	0.6284	0.6855
Популярность	Возраст	0.5697	0.6288	0.5781	0.6025	0.6004	0.6106	0.6338	0.6858
	Все				0.6268	0.6253	0.6134	0.6405	0.7104
LASSO	Возраст	0.5906	0.6099	0.5755	0.6001	0.5988	0.6014	0.5880	0.6412
	Все				0.6170	0.6159	0.5958	0.6040	0.6747
Среднее значение AUC внутри метода		0.5819	**0.6221**	0.5792	0.6099	0.6086	0.6135	**0.6197**	**0.6774**
Небалансированные выборки									
Альтман и Сабато	Возраст	0.5849	0.6064	0.5893	0.5828	0.5841	0.5443	0.6279	0.6538
	Все				0.6179	0.6197	0.5402	0.6282	0.6829
Популярность	Возраст	0.5604	0.5634	0.5619	0.5690	0.5680	0.5657	0.6302	0.6774
	Все				0.6027	0.6024	0.5415	0.6331	0.7076
LASSO	Возраст	0.6054	0.6006	0.5963	0.5950	0.5975	0.5266	0.5938	0.6134
	Все				0.6214	0.6229	0.5303	0.6065	0.6517
Среднее значение AUC внутри метода		0.5836	0.5901	0.5825	0.5981	**0.5991**	0.5414	**0.6200**	**0.6645**
Среднее значение AUC внутри метода по всем выборкам		0.5827	**0.6061**	0.5809	0.6040	0.6038	0.5775	**0.6199**	**0.6709**

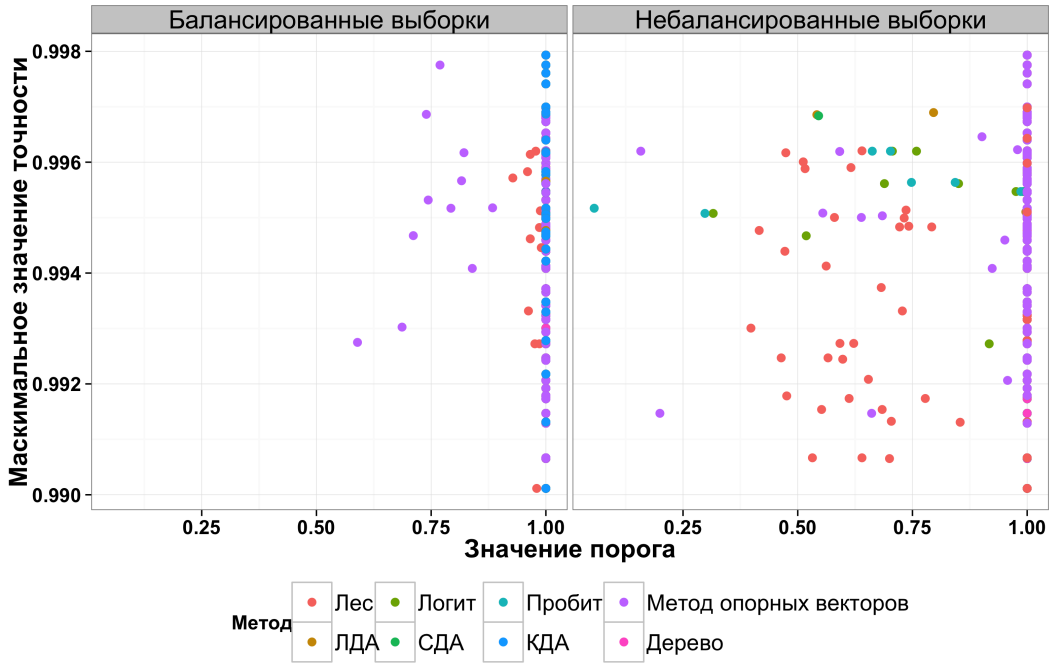
Для балансированных выборок по среднему значению AUC лидерами являются случайный лес, дерево и квадратичный дискриминантный анализ, а для небалансированных выборок — случайный лес, дерево и пробит-модель.

Обратимся к другим критериям сравнения кроме AUC. Как было сказано, критерий точности имеет большой недостаток. Продемонстрируем это на примере. На графике 21 отражено максимальное значение точности для каждого из построенных прогнозов в зависимости от значения порога, при котором оно достигалось.

Максимальное значение точности колеблется от 0.99 до 0.998. Одна-

ко если посмотреть на значения порога, которые соответствуют данным значениям точности, то большинство из них лежит около единицы. А по определению ROC-кривой порог, равный единице, означает, что все предприятия классифицированы как активные. Это можно объяснить тем, что в тестовых выборках было неравное число компаний-банкротов и активных компаний: количество активных компаний в разы превышало количество компаний-банкротов. По этой причине этот показатель не может адекватно характеризовать качество прогнозов.

Рис. 21: Максимальная точность и соответствующий ей порог



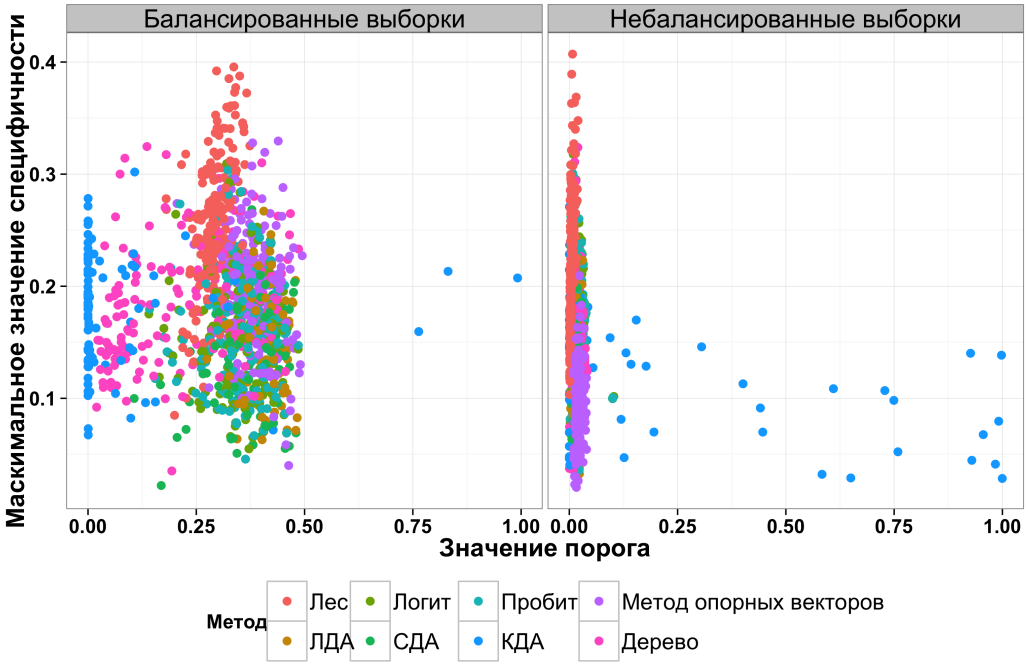
Рассмотрим ещё один критерий сравнения — специфичность при уровне чувствительности 0.9. Такой высокий уровень чувствительности задан, потому что мы хотим добиться верной классификации предприятий-банкротов с вероятностью 0.9. И при этой заданной вероятности оптимизируется уровень специфичности, то есть вероятность того, что активное предприятие будет классифицировано как активное, а не как банкрот.

На графике 22 отображён максимальный уровень специфичности при уровне чувствительности равном 0.9, в зависимости от значения порога,

типа выборки и метода.

Пороги определяются прежде всего внутренним устройством метода. На графике 22 видно, что пороги, при которых достигается наибольшая специфичность для балансированной выборки, в целом существенно выше, чем для небалансированной.

Рис. 22: Максимальная специфичность при чувствительности 0.9



Связано это со следующим эффектом. Условно выборку можно разделить на две части: «легко» классифицируемые предприятия, у которых по объясняющим переменным легко прогнозируется их статус, и «трудно» классифицируемые предприятия. «Трудно» классифицируемые предприятия составляют существенную часть выборки, и для них вероятность банкротства прогнозируется близкой к доле банкротств по выборке. Существенное количество прогнозных вероятностей близких к доле банкротств по выборке влияет на значения порога.

При использовании тривиального метода «каждое предприятие считается банкротом с вероятностью 0.9» достигается чувствительность 0.9 и

специфичность 0.1. Среди рассмотренных методов максимальная специфичность достигает 0.4, что говорит о значительном улучшении качества прогноза по сравнению с тривиальной моделью.

По максимальной специфичности при чувствительности 0.9 случайный лес оказался опять же лучше, чем остальные методы.

В приложении приведены графики максимального значения специфичности при чувствительности 0.9 в зависимости от метода и далее от типа формулы. Эти графики аналогичны графикам AUC, которые приведены раньше в работе. Они приведены в приложении, потому что выводы по ним такие же, как и по AUC: случайный лес опережает все методы и на баласированных выборках, и на небалансированных выборках. Медианы оставшихся методов довольно близки между собой, хотя на небалансированных выборках метод опорных векторов оказался худшим с наименьшим значением медианы.

9 Пример 2012 года

Рассмотрим модели на примере 2012 года. Будет использоваться отрасль оптовой и розничной торговли и балансированная выборка. Это отрасль выбрана, так как среди исследуемых секторов экономики она самая многочисленная.

9.1 Интерпретация коэффициентов в логит-моделях

На следующей странице приведены коэффициенты в логит-моделях, оценённых по шести формулам. Сначала рассмотрим влияние финансовых переменных на вероятность быть банкротом. На вероятность стать банкротом влияют показатели рентабельности, ликвидности и финансового рычага.

Показатель рентабельности roa (Net income / Total assets) значим при уровне значимости 10% во всех четырёх моделях, где он присутствует. С ростом рентабельности вероятность стать компанией-банкротом падает, потому что увеличение этого отношения означает, что компания получает больший доход и может в случае необходимости пустить большую часть на выплату долга.

Показатели ликвидности liq (Cash / Total assets) и sr (Total equity / Total assets) значимы при уровне значимости 1% и 0.1% соответственно.

Рост первого показателя ликвидности (liq) означает, что вероятность дефолта увеличивается. Это можно объяснить тем, что компании, которые держат избыточную наличность, а не реинвестируют в новые разработки, вскоре перестают расти и начинают испытывать финансовые сложности.

Таблица 12: Логит-модели для отрасли оптовой и розничной торговли в 2012 году

	Альтман + Возраст	*Альтман + Все*	*Популярные + Возраст*	*Популярные + Все*	*LASSO + Возраст*	*LASSO + Все*
Intercept	0.150 (0.056)**	0.642 (0.248)**	0.216 (0.050)***	1.101 (0.252)***	0.426 (0.064)***	0.549 (0.299)·
iptd	0.693 (0.353)*	0.630 (0.327)·	0.022 (0.031)	0.026 (0.043)		
ebta	−0.016 (0.020)	−0.020 (0.021)	−0.009 (0.009)	−0.008 (0.009)		
stdte	0.000 (0.000)*	0.000 (0.000)*				
roa	−0.039 (0.022)·	−0.038 (0.021)·	−0.007 (0.004)·	−0.008 (0.004)·		
liq	0.472 (0.144)**	0.580 (0.149)***				
sr	−0.162 (0.040)***	−0.184 (0.041)***				
cr			0.002 (0.001)	0.002 (0.001)		
wcta			0.008 (0.005)	0.008 (0.005)·	0.029 (0.072)	−0.007 (0.076)
lr			0.001 (0.002)	0.001 (0.002)		
tdta			0.007 (0.005)	0.008 (0.005)		
gg					−0.212 (0.032)***	−0.240 (0.034)***
nat					0.000 (0.000)	0.000 (0.000)
ebtm					−0.425 (0.175)*	−0.351 (0.166)*
ltdta					−1.331 (0.644)*	−1.713 (0.687)*
Age	−0.040 (0.007)***	−0.042 (0.008)***	−0.048 (0.007)***	−0.050 (0.008)***	−0.057 (0.009)***	−0.056 (0.009)***
Micro		−1.090 (0.135)***		−1.228 (0.139)***		−1.591 (0.165)***
Small		−0.965 (0.135)***		−1.129 (0.138)***		−1.378 (0.161)***
Limited liability company		0.013 (0.210)		−0.272 (0.212)		0.779 (0.265)**
Far Eastern federal region		0.706 (0.184)***		0.329 (0.170)·		0.404 (0.190)*
North Caucasian federal region		1.074 (0.250)***		1.137 (0.257)***		1.133 (0.288)***
Northwest federal region		0.702 (0.113)***		0.707 (0.111)***		0.571 (0.128)***
Siberian federal region		1.024 (0.102)***		1.150 (0.105)***		1.105 (0.114)***
South federal region		0.328 (0.143)*		0.409 (0.148)**		0.528 (0.165)**
Ural federal region		0.389 (0.136)**		0.434 (0.135)**		0.225 (0.148)
Volga federal region		0.275 (0.104)**		0.413 (0.104)***		0.510 (0.121)***
AIC	5042.551	4858.187	5025.012	4811.855	4092.390	3888.233
BIC	5092.284	4970.087	5080.894	4929.827	4134.522	3990.555
Log Likelihood	−2513.275	−2411.094	−2503.506	−2386.928	−2039.195	−1927.117
Deviance	5026.551	4822.187	5007.012	4773.855	4078.390	3854.233
Num. obs.	3702	3702	3674	3674	3038	3038

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Рост второго показателя (sr) отрицательно влияет на вероятность банкротства, потому что, если капитал близок к величине суммарных активов, величина долга компании небольшая. Это, в свою очередь, означает, что она более способна расплачиваться по своим долгам. Также данный показатель часто интерпретируют как то, что получит каждый из собственников в случае банкротства предприятия. Таким образом, если эта величина растёт, то компания более устойчива в данный момент и вероятность банкроства соответственно ниже.

Показатель финансового рычага gg ((Non current liabilities + Loans) / Total equity) значим при любом разумном уровне значимости, а показатель ltdta (Long-term debt / Total assets) значим на 5%. С увеличением обоих из них вероятность быть банкротом падает.

Первый из них (gg) является аналогом отношения суммарного долга к капиталу, которое означает, что при привлечении дополнительного долга компания может осуществить те проекты, которые невозможно профинансировать только за счёт собственного капитала.

Второй показатель (ltdta) означает, что с ростом величины долгосрочных обязательств назначаются более низкие ежегодные или ежемесячные платежи, что повышает вероятность покрыть свои обязательства в сроки. Таким образом, вероятность дефолта уменьшается.

Теперь обратимся к нефинансовым показателям. Возраст, который присутстует во всех шести моделях, оказался значимым при любом разумном уровне значимости. С ростом возраста вероятность стать банкротом падает. Это подтверждает мнение о том, что новые предприятия чаще закрываются из-за неустойчивости функционирования, чем уже давно существующие компании.

Рассмотрим три модели, где кроме возраста есть ещё организационная форма, размер и федеральный округ. Дамми-переменные на размер тоже значимы на любом уровне значимости. За базовую категорию взяты компании среднего размера. При переходе компании из группы компаний среднего размера в группу компаний микро размера или малого размера вероятность испытать серьёзные финансовые сложности падает.

Дамми-переменная на вид организационной формы оказалась значима в двух из трёх моделей, хотя в третьей модели, где выбор переменных осуществлялся с помощью LASSO, она значима на уровне значимости 1%.

Перейдём к федеральным округам. Базовой категорией задан Центральный округ. Практически все дамми на округа значимы на любом разумном уровне значимости.

Таким образом, можно подытожить, что нефинансовые переменные необходимо включать в модель. Более того, если попарно сравнить модели с одинаковым набором финансовых переменных, но разным набором нефинансовых характеристик, с помощью таких показателей, как AIC, BIC, то модели с более полным набором нефинансовых переменных лучше.

9.2 Предельные эффекты в логит-моделях

Также стоит показать предельные эффекты в приведённых выше логит-моделях по оптовой и розничной торговле в 2012 году. Существует два способа рассчитывать предельные эффекты. Можно считать предельный эффект для средних значений всех регрессоров, а можно рассчитывать его как среднее арифметическое предельных эффектов по всем наблюде-

Таблица 13: Предельные эффекты в логит-моделях в 2012 году

	Альтман Возраст	*Альтман Все*	*Популярные Возраст*	*Популярные Все*	*LASSO Возраст*	*LASSO Все*
Intercept	0.037**	0.148*	0.053***	0.253***	0.102***	0.122
	(0.013)	(0.057)	(0.012)	(0.063)	(0.016)	(0.075)
iptd	0.169**	0.145**	0.005	0.006		
	(0.056)	(0.056)	(0.008)	(0.011)		
ebta	−0.004	−0.005	−0.002	−0.002		
	(0.004)	(0.005)	(0.002)	(0.002)		
stdte	0.000*	0.000*				
	(0.000)	(0.000)				
roa	−0.009*	−0.009·	−0.002·	−0.002·		
	(0.005)	(0.005)	(0.001)	(0.001)		
liq	0.115**	0.133***				
	(0.034)	(0.036)				
sr	−0.039***	−0.042***				
	(0.010)	(0.010)				
cr			0.000	0.000		
			(0.000)	(0.000)		
wcta			0.002	0.002	0.007	−0.001
			(0.001)	(0.001)	(0.018)	(0.019)
lr			0.000	0.000		
			(0.000)	(0.000)		
tdta			0.002	0.002		
			(0.001)	(0.001)		
gg					−0.051***	−0.053***
					(0.008)	(0.008)
nat					0.000	0.000
					(0.000)	(0.000)
ebtm					−0.102*	−0.078·
					(0.044)	(0.042)
ltdta					−0.319*	−0.381*
					(0.161)	(0.172)
Age	−0.010***	−0.010***	−0.012***	−0.012***	−0.014***	−0.013***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Micro		−0.243***		−0.298***		−0.375***
		(0.039)		(0.032)		(0.035)
Small		−0.222***		−0.259***		−0.307***
		(0.040)		(0.034)		(0.040)
Limited liability company		0.003		−0.062		0.173**
		(0.046)		(0.053)		(0.066)
Far Eastern federal region		0.136***		0.081*		0.100*
		(0.037)		(0.041)		(0.046)
North Caucasian federal region		0.188***		0.254***		0.258***
		(0.046)		(0.046)		(0.054)
Northwest federal region		0.139***		0.171***		0.140***
		(0.029)		(0.025)		(0.030)
Siberian federal region		0.197***		0.268***		0.263***
		(0.035)		(0.022)		(0.025)
South federal region		0.069*		0.100**		0.129**
		(0.030)		(0.035)		(0.039)
Ural federal region		0.081**		0.106**		0.056
		(0.029)		(0.032)		(0.037)
Volga federal region		0.059**		0.102***		0.126***
		(0.023)		(0.025)		(0.029)
AIC	5042.551	4858.187	5025.012	4811.855	4092.390	3888.233
BIC	5092.284	4970.087	5080.894	4929.827	4134.522	3990.555
Log Likelihood	−2513.275	−2411.094	−2503.506	−2386.928	−2039.195	−1927.117
Deviance	5026.551	4822.187	5007.012	4773.855	4078.390	3854.233
Num. obs.	3702	3702	3674	3674	3038	3038

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

ниям. В данном случае приведён второй вариант предельного эффекта, однако нужно отметить, что здесь он несильно отличается от первого варианта. Предельные эффекты для пробит-моделей практические не отличаются.

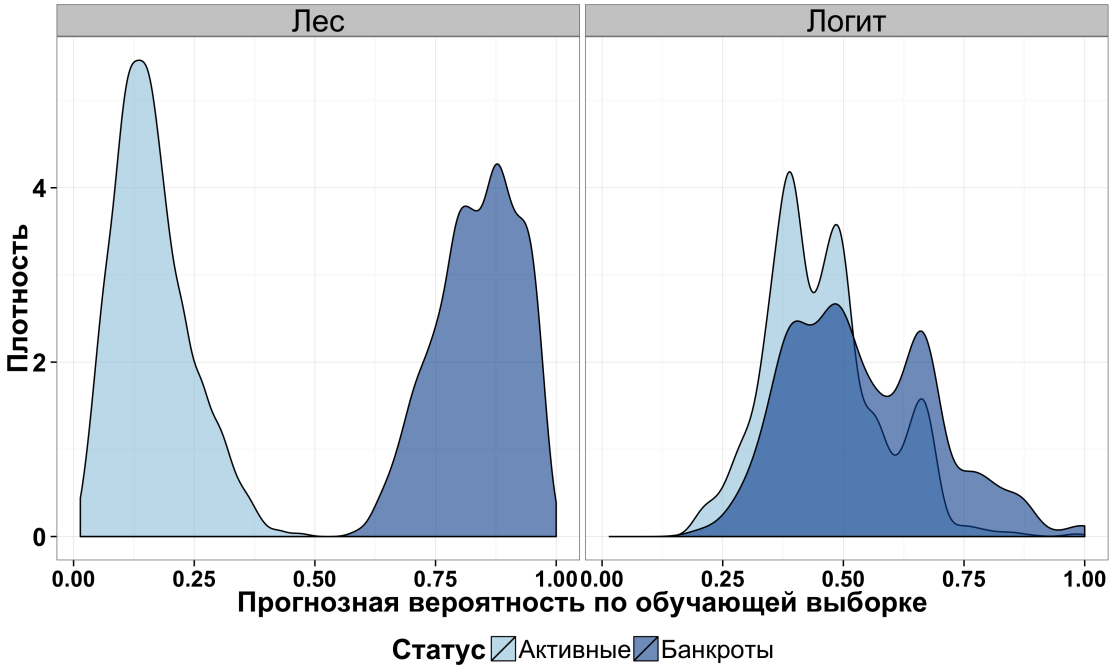
В модели Альтмана наиболее существенный вклад вносят iptd и liq. В модели с коэффициентами, отобранными по LASSO, наиболее существенной оказывается переменная ltdta, увеличиение ltdta на единицу приводит к уменьшению вероятности банкротства примерно на 0.3. Для большинства предприятий ltdta равен 0, а если предприятие имеет высокую долю долгосрочного долга, то это является индикатором его финансовой стабильности.

9.3 Сравнение алгоритма случайного леса и логит-модели

Теперь сравним случайный лес и логит-модель для отрасли оптовой и розничной торговли в 2012 году. Для примера возьмём модель, где есть все нефинансовые переменные и финансовые переменные, отобранные по частоте упоминаний в исследованиях. Выбрана именно эта модель, потому что ранее было показано, что для всех лет значения AUC для этой формулы были самыми высокими.

На графике 23 представлено распределение прогнозных вероятностей по модели для 2012 года для банкротов и активных предприятий. Прогнозы строились по той же выборке, по которой оценивалась модель. Поэтому они выглядят более «оптимистичными» по сравнению с прогнозами предыдущего раздела, где прогнозирование осуществляется на год вперёд.

Рис. 23: Сравнение алгоритма случайного леса и логит-модели

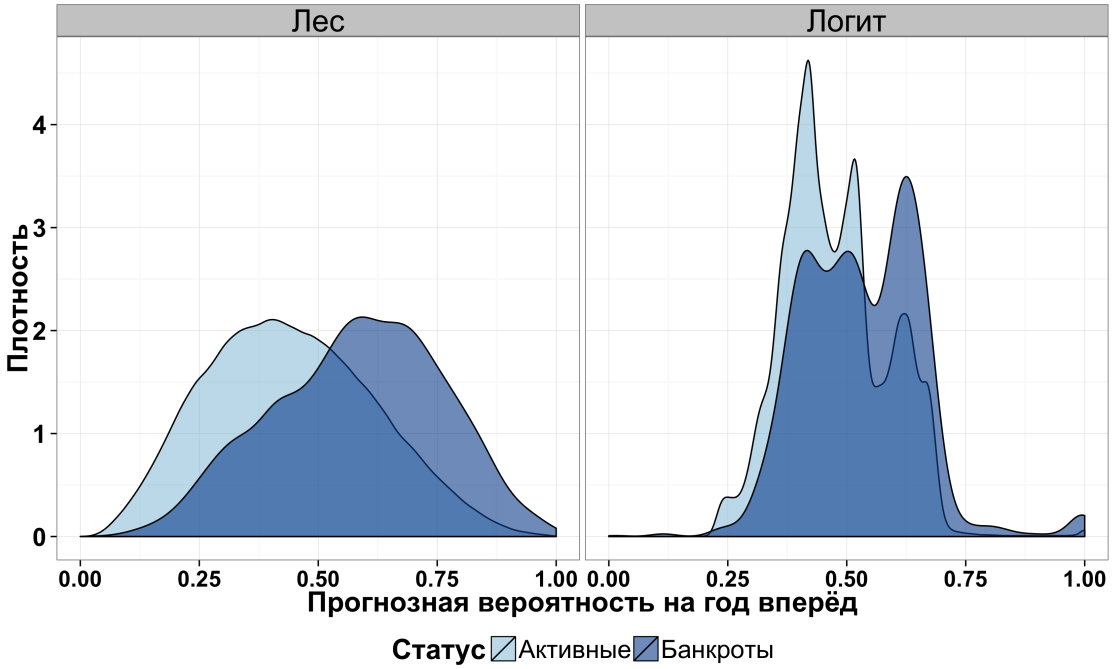


По графику видно, что лес значительно превосходит логит-модель: спрогнозированная вероятность для активных на самом деле компаний у леса не превышает 0.5, а для компаний-банкротов она всегда больше 0.5. Причём важно отметить, что для большинства активных компаний лежит около 0.1, а для компаний-банкротов — около 0.9. В то же время для логита картина абсолютна иная: этот метод зачастую классифицирует банкротов как активные предприятия и наоборот.

На следующей странице представлен график 24. Теперь модель была построена по данным 2011 года и честно спрогнозирована на данных 2012 года.

Здесь видно, что как логит-модель, так и лес неверно классифицируют приличную часть компаний обоих типов. Распределение прогнозных вероятностей для леса существенно более похоже на нормальное, чем у логит-модели: среднее прогнозной вероятности для активных компаний лежит около 0.35, а для дефолтных компаний — около 0.65.

Рис. 24: Сравнение алгоритма случайного леса и логит-модели



Если сравнивать график 23 и график 24, то заметна значительная разница в том, как случайный лес разделяет компании на той же выборке и на выборке на год вперёд. На той же выборке лес безошибочно разделил предприятия-банкроты от активных предприятий. В то же время прогноз на год вперёд оказался гораздо хуже, зато он отображает сложность поставленной задачи и даже в этом случае неверно классифицирует меньшее число наблюдений, чем логит-модель. Однако логит-модель прогнозирует приблизительно одинаково, что на «корректной» выборке, что на «некорректной выборке».

9.4 Важность переменных в алгоритме случайного леса

Возьмём ту же модель, оценённую с помощью алгоритма случайного леса, что и в предыдущем разделе, где сравнивались логит-модель и случайный лес.

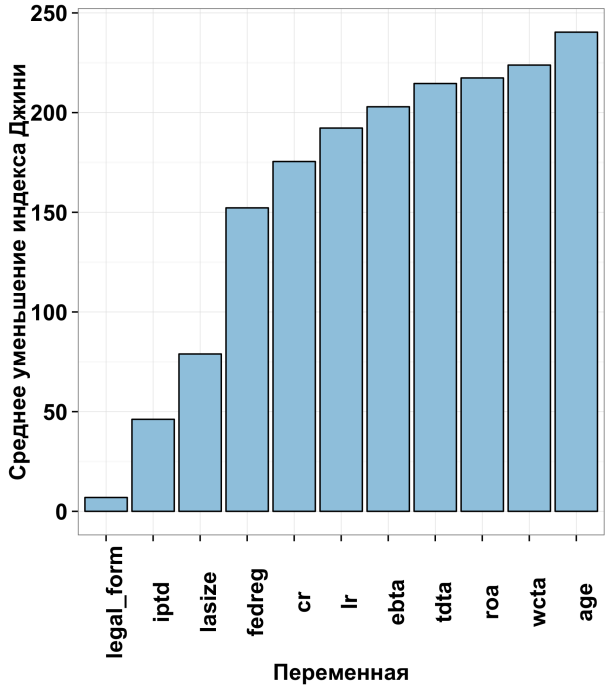
Для данного алгоритма не существует проверки гипотез на значимость коэффициентов. Однако аналогом может служить критерий сред-

него уменьшения индекса Джини по всем построенным деревьям, то есть степень важности каждой переменной в модели.

Допустим, что на некотором дереве узел с n предприятиями с долей банкротов h был разделен на два узла с n_1 и n_2 предприятиями и долями банкротов h_1 и h_2 соответственно. Падение индекса Джини в произвольном узле определяется как разница между индексом Джини самого узла, $2h(1-h)$, и индексом Джини после деления узла на две части, $\frac{n_1}{n}2h_1(1-h_1)+\frac{n_2}{n}2h_2(1-h_2)$.

Суммарное падение индекса Джини по всем узлам данного дерева, где происходит деление по переменной x, следовательно, может быть существенно больше единицы. В качестве показателя важности используют усреднённое по всем деревьям суммарное падение индекса Джини.

Рис. 25: Важность переменных в алгоритме случайного леса



На графике 25 представлены переменные из модели со всеми нефинансовыми показателями и финансовыми переменными, отобранными по степени их упоминаемости. По этой причине можно посмотреть, какие

именно из типов переменных вносят наибольший вклад в определение исхода.

Для начала стоит сказать, что значения для организационной формы настолько мало, что если бы мы говорили в терминах проверки гипотезы о значимости коэффициента, то этот коэффициент был бы незначим. Ко второй группе переменных, оказывающих несильное влияние на исход, относятся последний доступный размер и iptd (Interest paid / Total debt). Оставшиеся переменные оказывают весомое влияние на вероятность банкротства. Среди них как нефинансовые, так и финансовые характеристики фирм. Причём важно отметить, что самым важным критерием для случайного леса оказался возраст компании.

9.5 Таблицы сопряжённости для алгоритма случайного леса

Построим таблицу сопряжённости для трёх значений порога, 0.4, 0.5 и 0.6 для описанной выше модели, оценённой с помощью алогоритма случайного леса. Как и следовало ожидать, с ростом порога чувствительность падает, а специфичность растёт.

Таблица 14: Матрицы сопряжённости для разных порогов

Порог	Прогнозируемый исход	Фактический исход		Чувствительность	Специфичность
		Положительный	Отрицательный		
0.4	Положительный Отрицательный	1 526 (TP) 311 (FN)	150 936 (FP) 115 915 (TN)	0.8307	0.4344
0.5	Положительный Отрицательный	1 255 (TP) 582 (FN)	97 050 (FP) 169 801(TN)	0.6832	0.6363
0.6	Положительный Отрицательный	884 (TP) 953 (FN)	51 945 (FP) 214 906 (TN)	0.4812	0.8053

Так как для банка или другой кредитующей организации важно не выдать кредит предприятию-банкроту, то наиболее адекватным из трёх

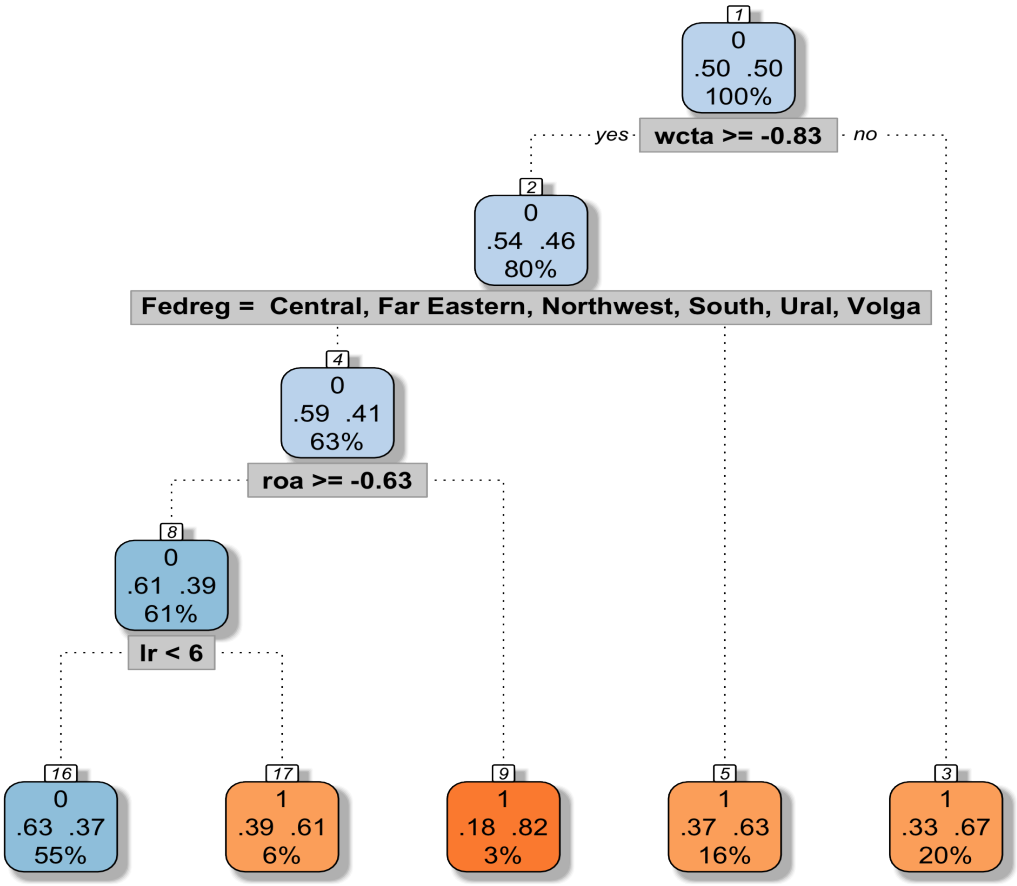
предъявленных представляется порог 0.4. Зависимость чувствительности и специфичности от порога симметрична относительно точки 0.5.

Для составления таблицы из трёх матриц неточностей модель оценивалась по данным 2011 года, а прогнозы делались по данным 2012 года. Если же использовать данные 2012 года и считать для них же прогнозные вероятности, то при пороге 0.5 данный случайный лес даст абсолютную точность равную единице.

9.6 Классификационное дерево

Для сравнения приведём такую же модель, как и в прошлом разделе, но оценённую с помощью метода классификационного дерева.

Рис. 26: Дерево (Активное — 0, Банкрот — 1)



Поясним некоторые обозначения на графике. Вверху каждого узла пишется, как были классифицированы предприятия данного узла: как активные (0) или как банкроты (1). В первом узле равное количество банкротов и активных, поэтому классификация произвольная. На второй строчке каждого узла пишется доля небанкротов и доля банкротов, а в последней строке — доля предприятий, попадающих в каждый узел.

Стоит отметить, что на этом графике ярко выражены различия между лесом и одним деревом. Здесь деление производится по wcta (Working capital / Total assets), потом по федеральному округу, далее по roa (Net income / Total assets) и по lr ((Current assets - stocks) / Current liabilities). В то же время возраста среди первых ветвей нет, и порядок двух следующих переменных обратный порядку важности переменных в модели, оценённой с помощью леса.

По графику 26 нельзя сделать вывод, какой из двух методов лучше, как это можно было видеть на графиках ранее. Однако такого вида дерево позволяет привести простые и ясные правила для классификации.

10 Заключение

Итак, проделав данное исследование, которое посвящено прогнозированию финансовых сложностей российских компаний среднего и малого бизнеса четырёх отраслей: оптовой и розничной торговли, обрабатывающих производств, строительства и операций с недвижимостью — можно сделать следующие выводы:

- Все финансовые переменные имеют распределения с тяжёлыми хвостами. По этой причине средние и стандартные отклонения очень изменчивы во времени. При этом робастные характеристики: медианы и медианные абсолютные отклонения довольно стабильны в течение всего исследуемого периода, даже в период кризиса.

- Среди используемых методов прогнозирования: логит- и пробит-моделей, линейного дискриминантного анализа, квадратичного дискриминантного анализа, дискриминантного анализа смеси распределений, метода опорных векторов, классификационного дерева и случайного леса — лучшим методом оказался случайный лес, что подтверждает тот факт, что зависимость между финансовыми, нефинансовыми переменными и вероятностью дефолта является нелинейной.

- Два выбранных критерия качества прогнозов: площадь под ROC-кривой и максимальная специфичность при чувствительности 0.9 — выделяют в целом одни и те же модели и методы.

- С помощью перечисленных выше методов было оценено два типа моделей: с финансовыми переменными и возрастом компании, а также

с финансовыми переменными и остальными нефинансовыми характеристиками компании, а не только возрастом. Финансовые переменные отбирались тремя способами. Первый способ — адаптировать модель Альтмана и Сабато для компаний среднего и малого бизнеса, подбирая аналогичные финансовые отношения. Второй способ — взять финансовые переменные, которые наиболее часто упоминались в исследованиях для прогнозирования банкротства. Третий способ — отобрать переменные с помощью LASSO. Второй способ выбора финансовых отношений и добавление нефинансовых характеристик оказался самым успешным.

- Добавление нефинансовых переменных таких, как федеральный округ размер компании, в любую модель улучшает её способность прогнозировать.

- Среди двух исследуемых организационных форм (ЗАО и ООО) не найдено существенного отличия во влиянии на вероятность банкротства.

- Возраст компании, который берётся в качестве объясняющей переменной во всех типах моделей, значим всегда. Увеличение возраста влияет отрицательно на вероятность банкротства, поэтому он тоже должен учитываться при прогнозировании.

- Среди финансовых переменных наиболее важными оказались показатели рентабельности, ликвидности и финансового рычага.

- Наличие в выборках, по которым строились модели, неравного количестве банкротов и активных компаний (активных компаний гораздо больше) не повлияло значительно на качество прогнозов по

сравнению моделями, которые строились по балансированным выборкам с равным числом компаний каждого типа.

- Были обнаружены структурные сдвиги, которые трудно идентифицировать по описательным статистикам переменных. В оптовой и розничной торговле сдвиг произошёл в 2010 году: все модели 2009 года существенно хуже предсказывают исходы на 2010 год. Вероятно, это связано с кризисом 2008 — 2009 годов. В строительной отрасли и обрабатывающих производствах подобные сдвиги были в 2007 и 2009 годах. Отрасль операций с недвижимостью стала менее изменчивой с 2009 года.

- Исследуемые отрасли существенно отличаются друг от друга как по влиянию отдельных переменных, так и и по структуре временной зависимости. Однако наилучшая модель для них одна и та же.

- Существенно отличается качество прогнозов по обучающей выборке и вне неё. При использовании случайного леса достигается 100%-ная точность в первом случае, в то время как при корректном прогнозе можно добиться чувствительности и специфичности около 0.7.

В дальнейшем можно распространить исследование на другие отрасли. Более того, можно применять другие методы оценивания моделей такие, как нейронные сети или бустинг. Основная проблема, с которой придётся столкнуться в любом случае, — заполнение пропущенных значений.

11 Список литературы

1. Altman, E. I., (1968), "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The journal of finance*, vol. 23, No. 4, pp. 589–609.

2. 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, pp. 505–529.

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

4. Beaver, W. H., (1966), "Financial ratios as predictors of failure", *Journal of accounting research*, vol. 4, pp. 71–111.

5. 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, pp. 5–36.

6. Ciampi, F., Vallini, C., Gordini, N., (2009), "Using Artificial Neural Networks Analysis for Small Enterprise Default Prediction Modeling: Statistical Evidence from Italian Firms", vol. 1, pp. 1–26.

7. 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.

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

9. Härdle, W. K. и др., (2007), "The default risk of firms examined with smooth support vector machines", *Discussion papers, German Institute for Economic Research*, No. 757, pp. 1–30.

10. 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.

11. Kapliński, O., (2008), "Usefulness and credibility of scoring methods in construction industry", *Journal of Civil Engineering and Management*, vol. 14, No. 1, pp. 21–28.

12. Khorasgani, A., (3 фев. 2014), "Optimal accounting based default prediction model for the UK SMEs".

13. 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, pp. 1–26.

14. 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, pp. 301–313.

15. 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, pp. 256–271.

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

17. Martin, D., (1977), "Early warning of bank failure: A logit regression approach", *Journal of Banking & Finance*, vol. 1, No. 3, pp. 249–276.

18. Ohlson, J. A., (1980), "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of accounting research*, vol. 18, No. 1, pp. 109–131.

19. 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, pp. 847–868.

20. 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, pp. 25–54.

21. 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.

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

23. Wei, L., Li, J., Chen, Z., (2007), "Credit risk evaluation using support vector machine with mixture of kernel", *Computational Science–ICCS*, pp. 431–438.

24. Wilson, R. L., Sharda, R., (1994), "Bankruptcy prediction using neural networks", *Decision support systems*, vol. 11, No. 5, pp. 545–557.

25. Zeitun, R., Tian, G., Keen, K., (2007), "Default probability for the Jordanian companies: A test of cash flow theory", vol. 8, pp. 147–162.

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

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

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

12 Приложение

12.1 Таблицы финансовых отношений

Таблица 15: Показатели финансового рычага

Формула в данных	**Название в данных**
Bank loans / Turnover	lt
Capital / Total Liabilities	ctl
(Current liabilities − Cash) / Total assets	clcta
Long-term debts / Total assets	ltdta
Non-current liabilities / Total assets	ltlta
Non-current liabilities / Total capital	ltltc
(Non current liabilities + Loans) / Shareholders' funds	gg
Short-term debt / Shareholders' funds	stdte
Shareholders' funds / Total liabilities	sol
Total capital / Fixed assets	tcfa
Total debts / Total assets	tdta
Total debts / (Total debts + Shareholders' funds)	tdtde
Total liabilities / Total capital	tltc
Total liabilities / Total assets	tlta

Таблица 16: Показатели ликвидности

Формула в данных	**Название в данных**
Cash / Current liabilities	ccl
Cash / Sales	cs
Cash / Total assets	liq
(Cash + Debtors) / Current liabilities	cdcl
(Cash + Debtors) / Total assets	cdta
Current assets / Current liabilities	cr
(Current assets − Stocks) / Current liabilities	lr
Current assets / Total liabilities	catl
Current liabilities / Total capital	cltc
Current assets / Sales	cas
Current liabilities / Total assets	clta
Fixed assets / Total assets	fata
Shareholders' funds / Non current liabilities	shlr
Shareholders' funds / Total assets	sr
Working capital / Total assets	wcta

Таблица 17: Показатели рентабельности

Формула в данных	**Название в данных**
Net Income / Sales	nisa
Net Income / Shareholders' funds	roe
Net Income / Total assets	roa
(Net income + interest paid) / (Shareholders' funds + Non-current liabilities)	roce
EBIT / Capital employed	ebce
EBIT / Sales	ebsa
EBIT / Total assets	ebta
EBIT / Total liabilities	ebtl
EBIT / Turnover	ebtm
EBITDA / Operating revenue	ebdor
EBITDA / Total assets	ebdta
EBITDA / Total liabilities	ebdtl
Gross profit / Sales	gps
Gross profit / Total assets	gpta
Gross profit / Operating revenue	gm
Profit before tax / Operating revenue	pm
Profit before tax / Shareholders' funds	roea
Profit before tax / Total assets	roaa
(Profit before tax + Interest paid) / (Shareholders' funds + Non-current liabilities)	rocea
Turnover / Number of employees	tne

Таблица 18: Показатели обслуживания долга

Формула в данных	**Название в данных**
Collection period / Credit period	cpcr
Cost of goods sold / Creditors	cgsc
(Creditors / Turnover) · 360	crpd
(Debtors / Turnover) · 360	clpd
Interest paid / Total debt	iptd
EBIT / Interest paid	ic
EBITDA / Interest paid	cov
Sales / Fixed assets	sfa
Sales / Debtors	sd
Turnover / (Shareholders' funds + Non-current liabilities)	nat
Turnover / Stocks	st

Таблица 19: Показатели активности

Формула в данных	**Название в данных**
Debtors / Total debts	dtd
Net Income / Turnover	act
Sales / Creditors	sc
Sales / Current Assets	sca
Sales / Total assets	sata
Total assets / Turnover	tat

Таблица 20: Описание некоторых финансовых показателей

Capital employed = Total assets − Current liabilities
Cash = Cash and Cash equivalent
EBIT = Operating profit or loss
EBITDA = Operating profit or loss + Depreciation
Net Income = Profit or loss for period
Shareholders' funds = Total Equity
Short-term debt = Creditors + Loans
Total capital = Total shareholders' funds and liabilities
Total debts = Short-term debt + Long-term debt
Total liabilities = Non-current liabilities + Current liabilities
Turnover = Operating revenue

12.2 Площадь под ROC-кривой для различных методов

Рис. 27: Площадь под ROC-кривой для линейного дискриминантного анализа

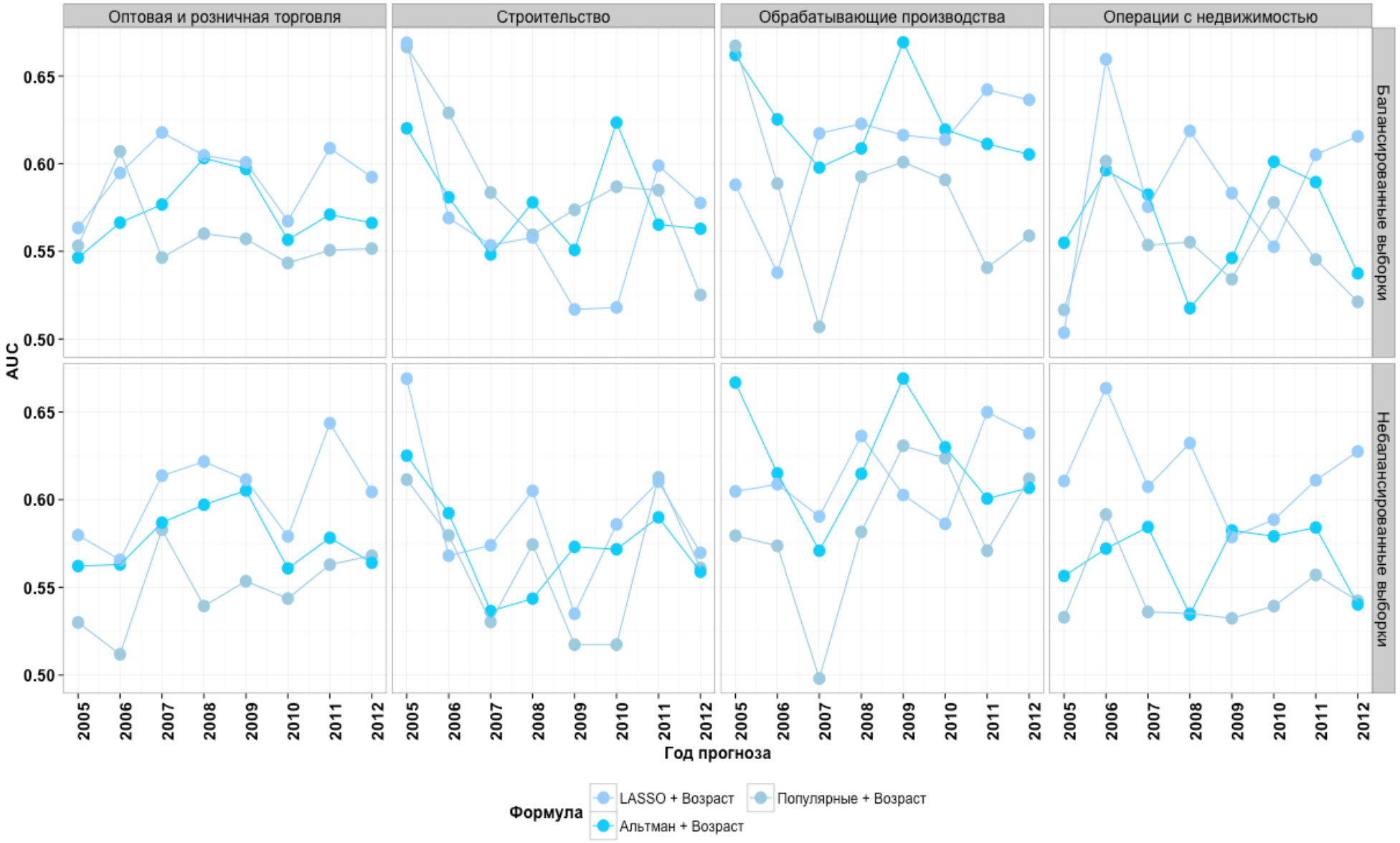


Рис. 28: Площадь под ROC-кривой для смешанного дискриминантного анализа

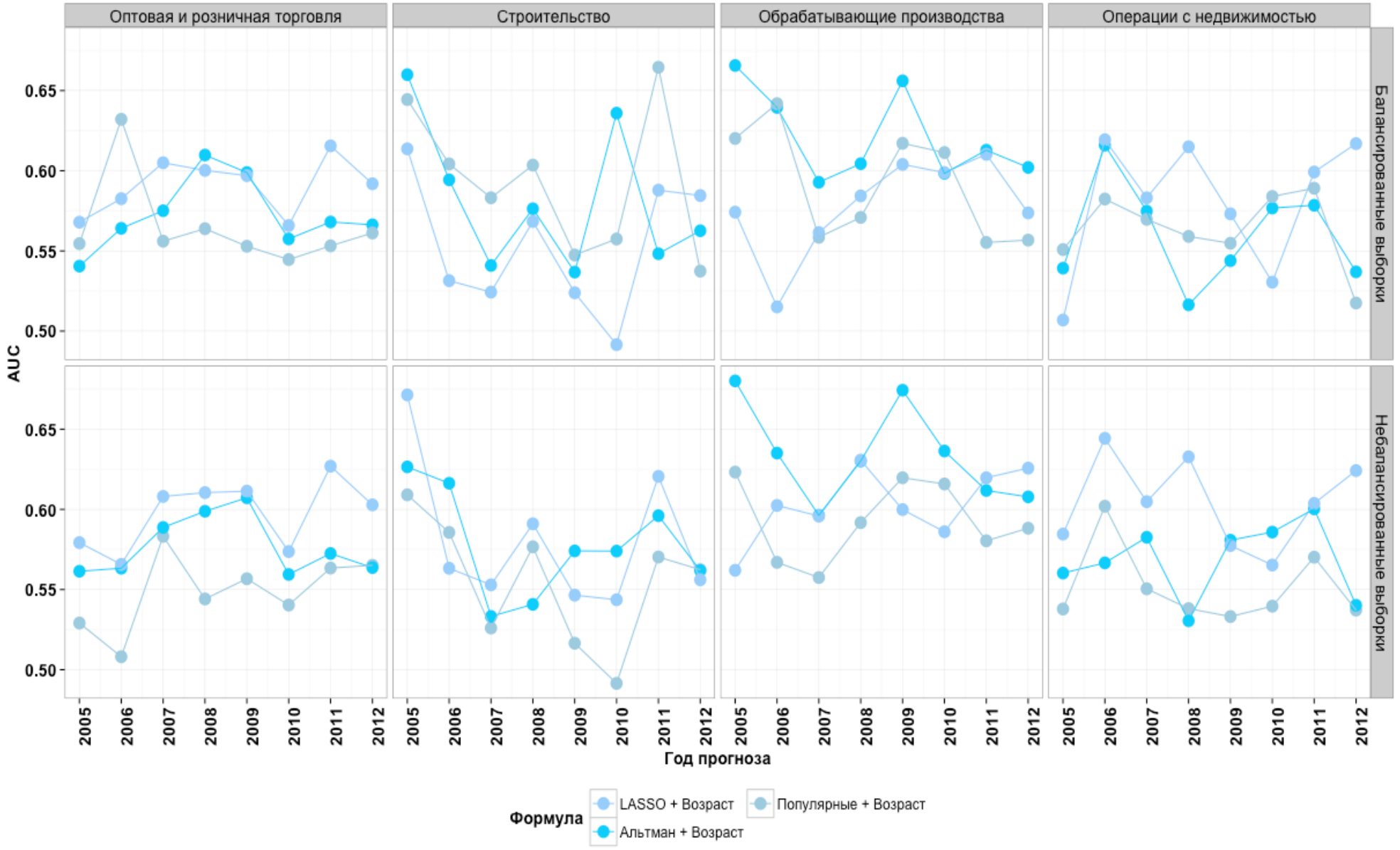


Рис. 29: Площадь под ROC-кривой для квадратичного дискриминантного анализа

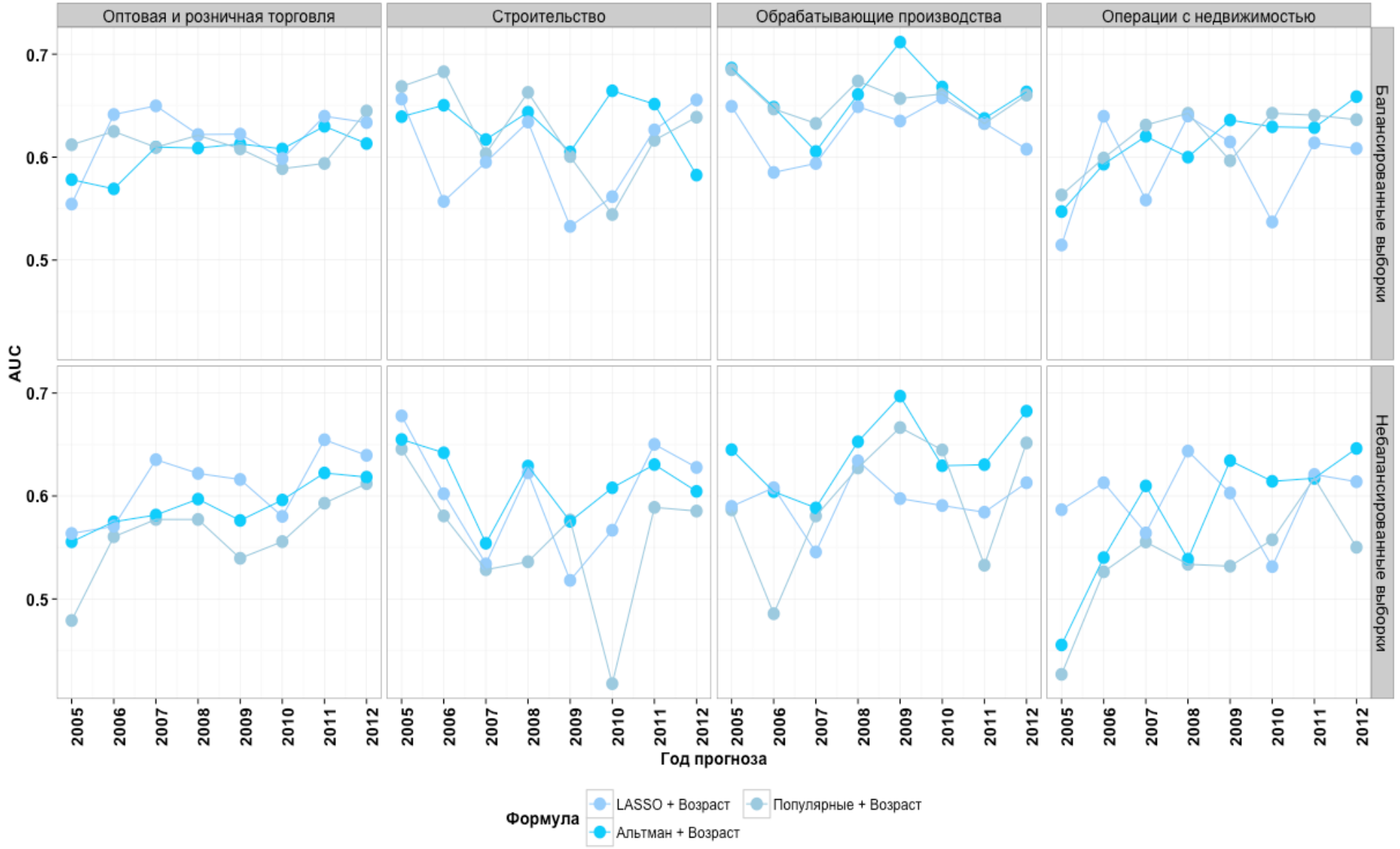


Рис. 30: Площадь под ROC-кривой для логит-модели

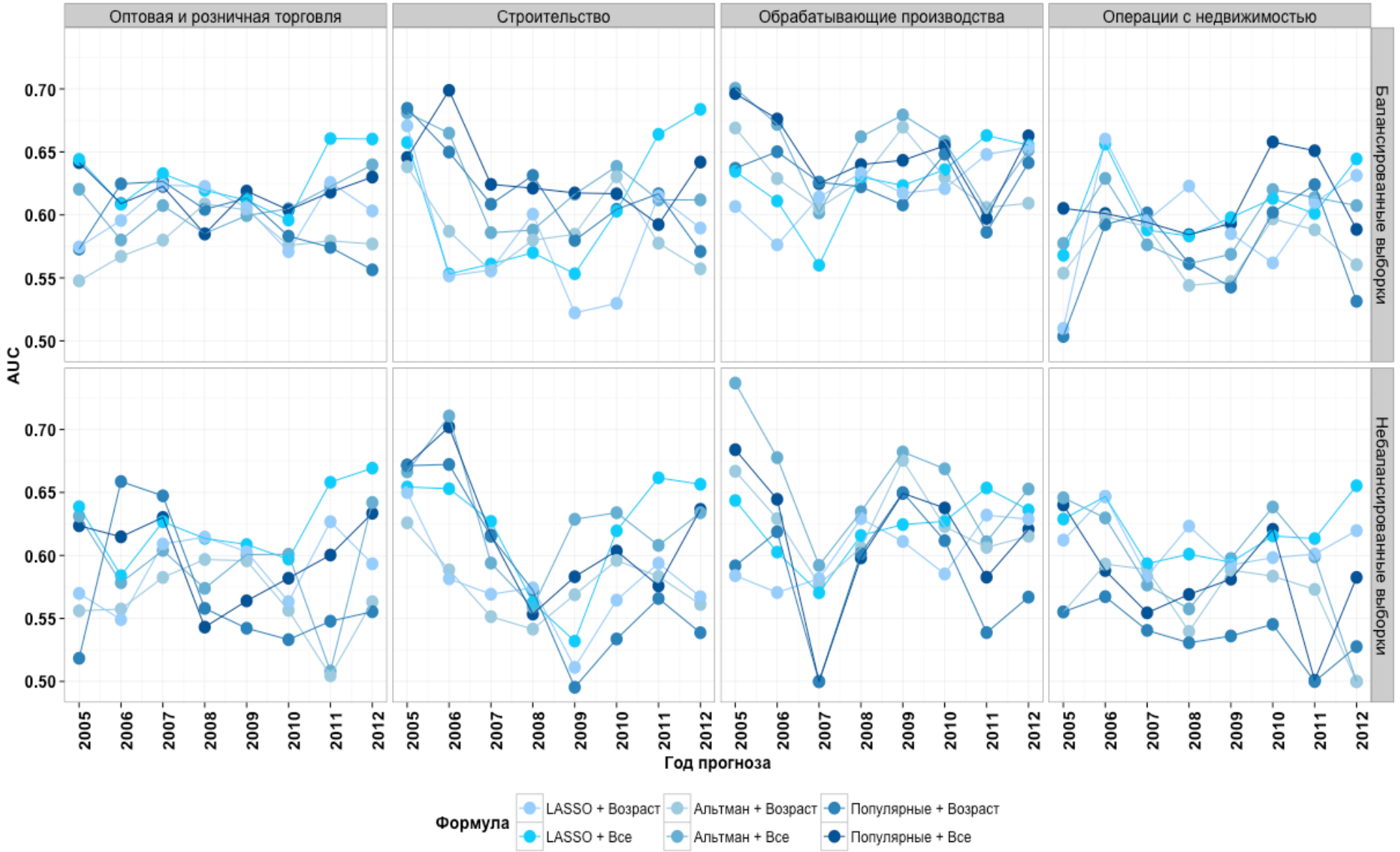


Рис. 31: Площадь под ROC-кривой для пробит-модели

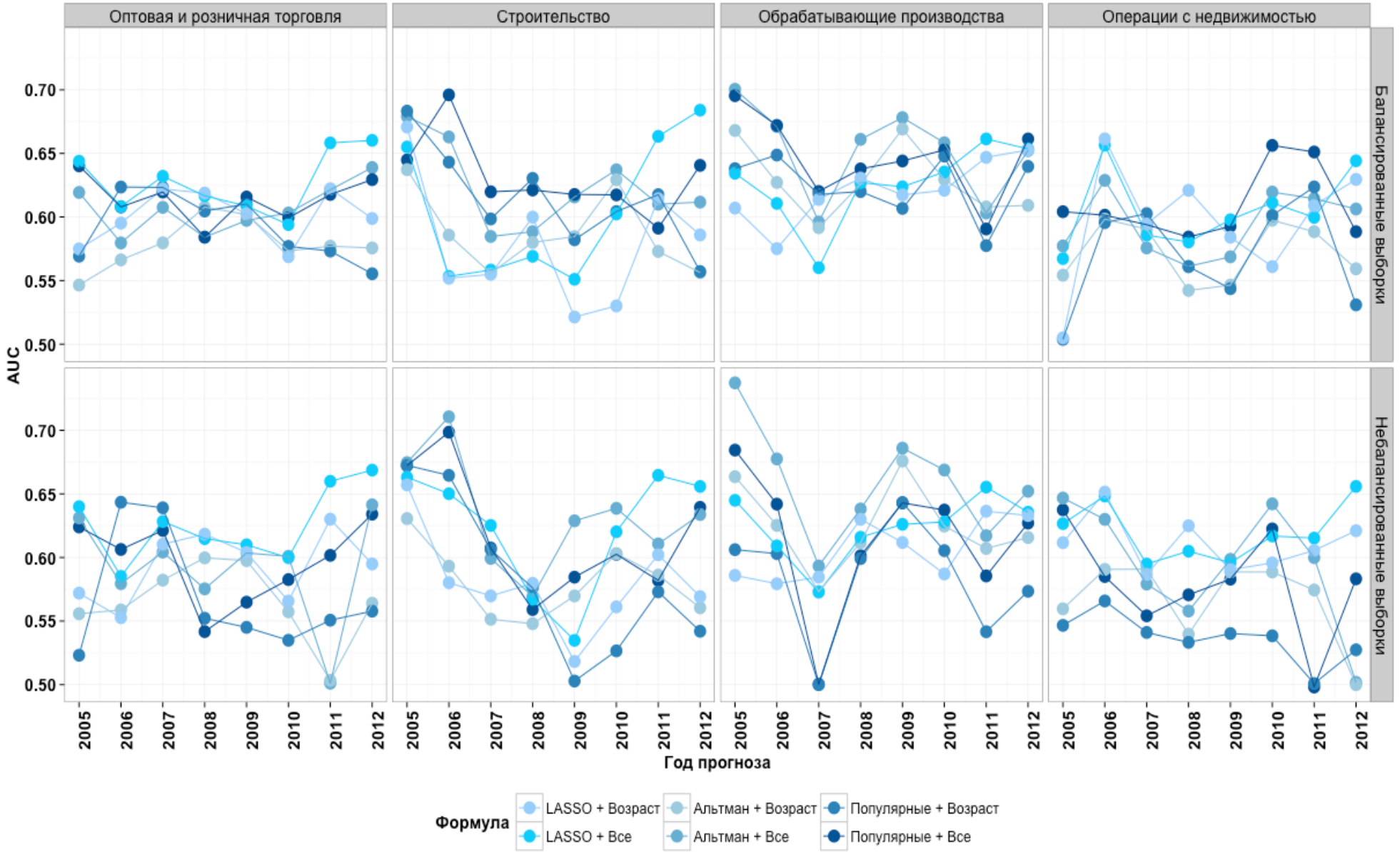


Рис. 32: Площадь под ROC-кривой для метода опорных векторов

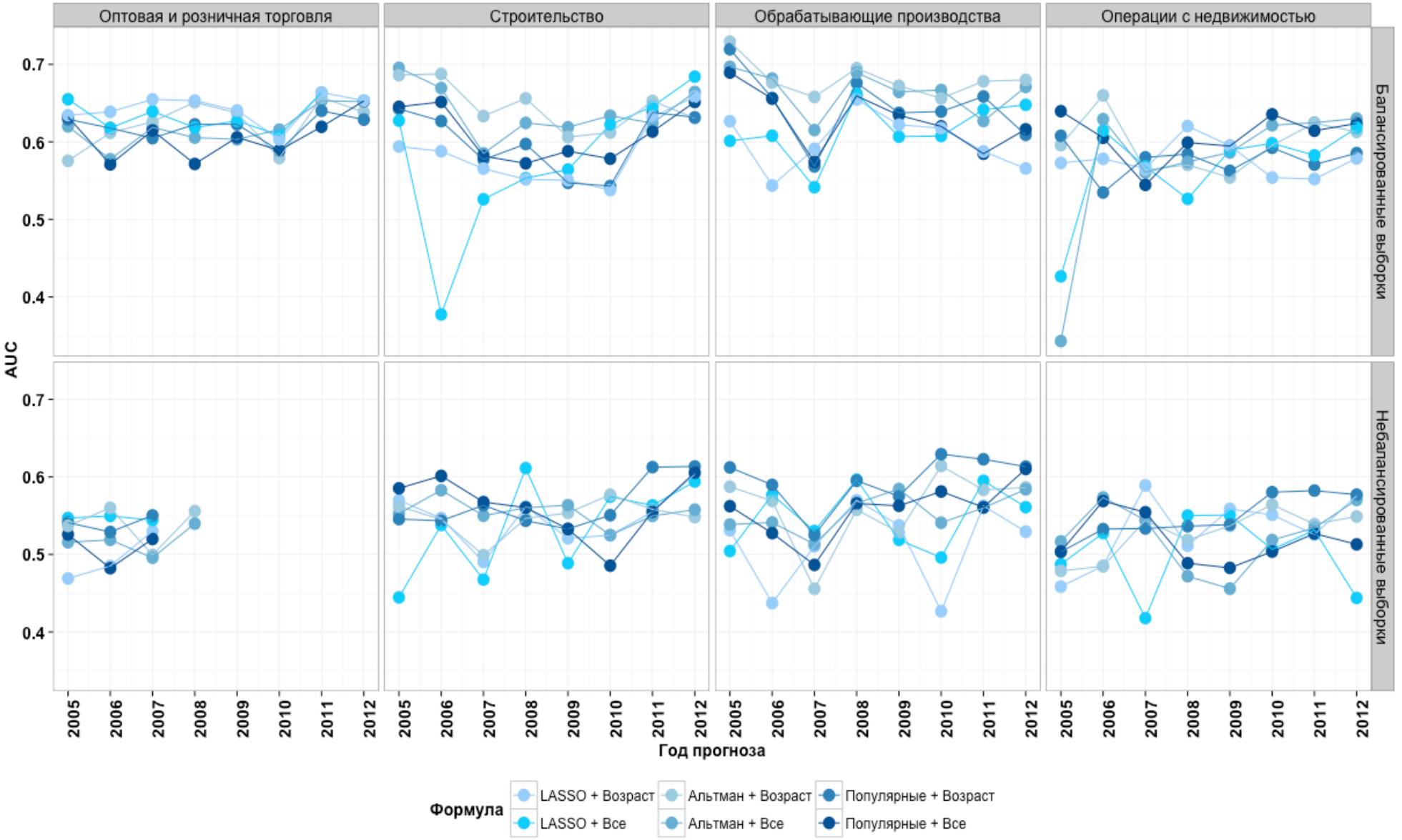
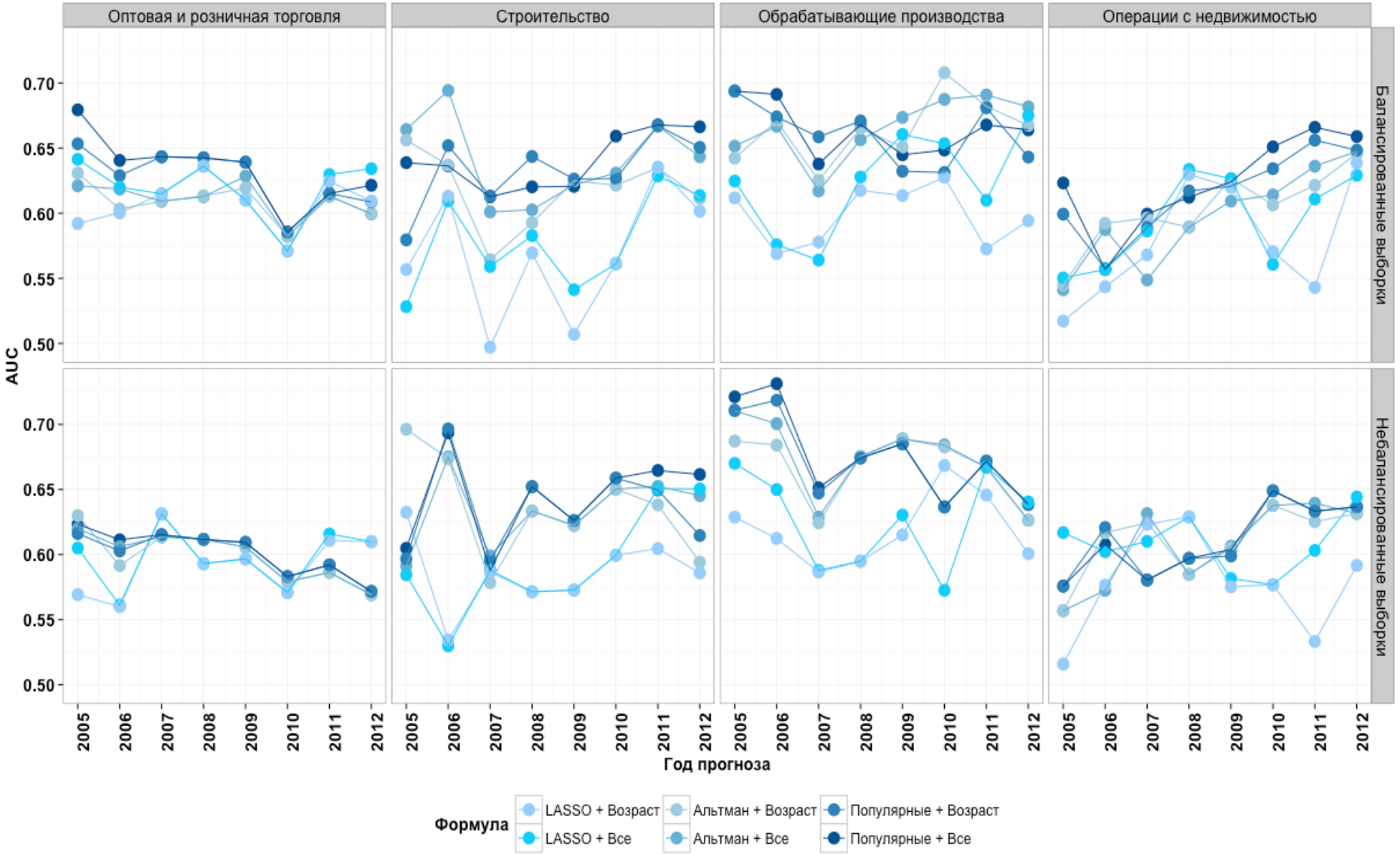


Рис. 33: Площадь под ROC-кривой для классификационного дерева



12.3 Специфичность при чувствтельности 0.9

Рис. 34: Распределение площади под ROC-кривой в зависимости от метода

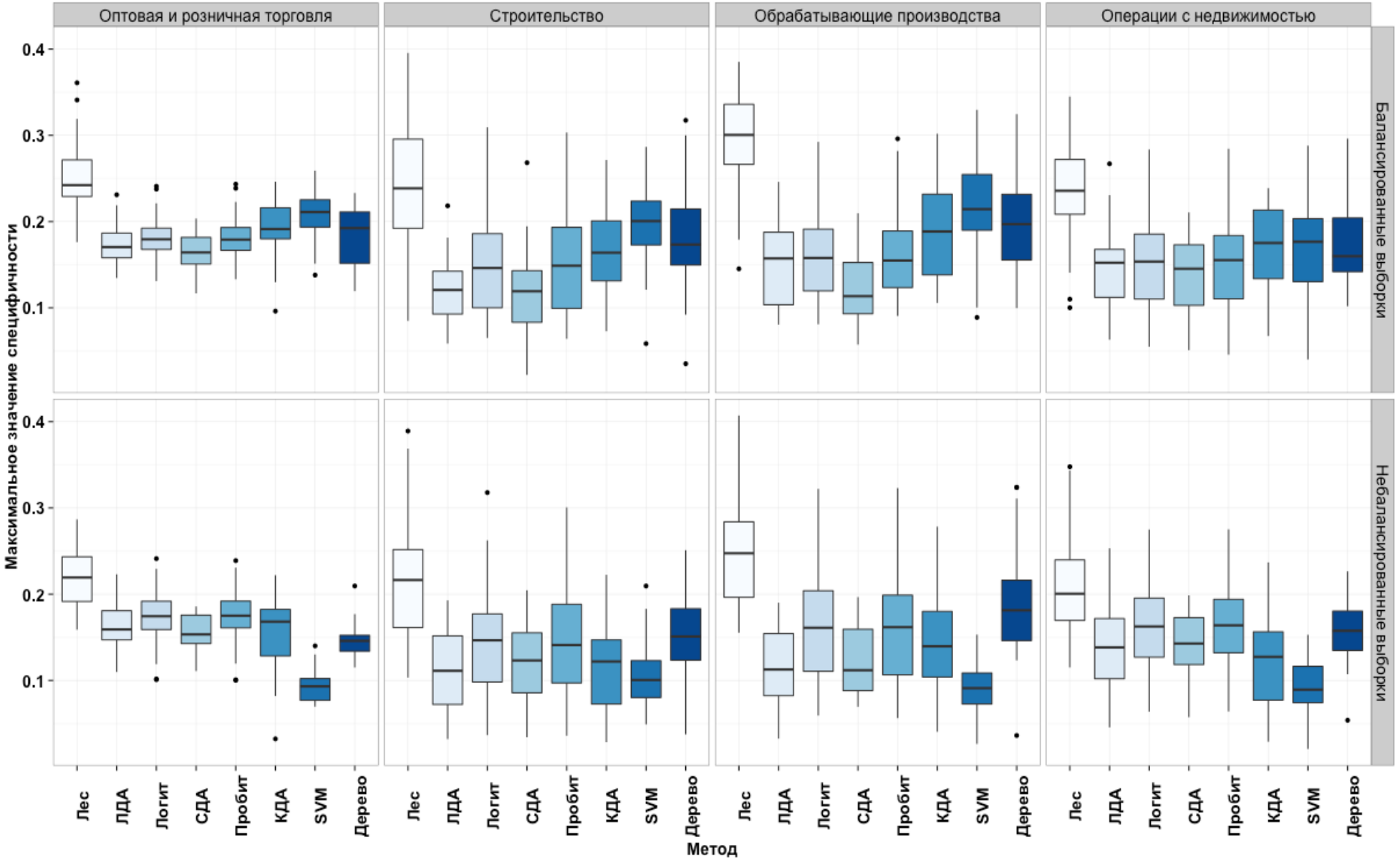


Рис. 35: Распределение площади под ROC-кривой в зависимости от формулы