Банк России

Модель вероятности дефолта с использованием транзакционных данных российских компаний

Серия докладов об экономических исследованиях

№ 97 / июнь 2022

А. Шевелев
Г. Бузанов

Андрей Шевелев
Банк России, Департамент исследований и прогнозирования
E-mail: shevelevaa@cbr.ru

Глеб Бузанов
Банк России, Департамент финансовой стабильности
E-mail: buzanovgs@cbr.ru

Авторы благодарят Сергея Селезнева, Сергея Глечикова и участников внутреннего исследовательского семинара Банка России за полезные комментарии и предложения. Все ошибки и упущения принадлежат авторам.

Статьи, выходящие в Серии докладов об экономических исследованиях, анонимно рецензируются членами Консультативного совета по научным исследованиям Банка России и внешними рецензентами.

Фото на обложке: Shutterstock/FOTODOM

107016, г. Москва, ул. Неглинная, 12
+7 (495) 771-91-00, +7 (495) 621-64-65 (факс)
Официальный сайт Банка России: www.cbr.ru

Оглавление

Резюме

Цель данного исследования – проверить полезность транзакционных данных платежной системы Банка России для прогнозирования вероятности дефолта российских компаний. Для достижения этой цели мы строим модели вероятности дефолта с использованием методов машинного обучения на основе данных годовой бухгалтерской отчетности по каждой отраслевой группе. Затем в модели мы добавляем признаки, созданные на основе транзакционных данных, что улучшает качество прогноза согласно метрике ROC AUC.

Кроме того, мы обучаем модели вероятности дефолта для каждой отраслевой группы с использованием алгоритма случайного леса (Random Forest) только на основе данных платежной системы Банка России. Качество такого прогноза в среднем несколько ниже согласно метрике ROC AUC, но эти оценки могут быть получены по крайней мере на три месяца раньше, чем оценки на основе данных годовой бухгалтерской отчетности.

Наши результаты подтверждают полезность транзакционных данных платежной системы Банка России для качества прогнозирования вероятности дефолта российских компаний. Кроме того, оценка важности признаков методом случайного леса показывает, что основными источниками дополнительной информации являются налоги на заработную плату и социальные выплаты.

Ключевые слова: модель вероятности дефолта, транзакционные данные, случайный лес, логистическая регрессия

Коды JEL: C53, C5, E44

Введение

Обеспечение бесперебойной работы финансового сектора и повышение его стабильности – одна из ключевых функций Банка России. В рамках этих задач осуществляется постоянный мониторинг финансовой системы и применяются инструменты макропруденциальной политики. Для этого Банк России учитывает в том числе оценки вероятностей дефолта российских компаний, содержащихся в портфелях коммерческих банков.

В условиях современной экономической системы банкам важно своевременно и точно оценивать вероятность дефолта компаний. Если прогнозы финансового состояния заемщика слишком консервативны, банк может классифицировать вполне надежных заемщиков как неплатежеспособных; следовательно, возникнет необходимость в создании избыточных резервов по кредитам, что может приводить к формированию у банков неоптимальных структур капитала и снижать объем кредитных ресурсов. Напротив, чрезмерный оптимизм может стать причиной дефицита резервов, что повлечет за собой риск дефолта банка и, как следствие, проблемы в финансовой системе в целом.

Большинство моделей прогнозирования дефолтов компаний основаны на данных бухгалтерского учета и на информации о кредитах; в некоторых работах учитываются макроэкономические и финансовые показатели. Недостатками таких подходов можно считать низкую периодичность публикации данных (годовая бухгалтерская отчетность), задержка их публикации (на три месяца и более), а также отсутствие учета взаимодействия между экономическими агентами. Решить эти проблемы позволяют данные платежной системы Банка России (ПС БР).

Платежная система Банка России является «системно значимой платежной системой, через которую реализуется денежно-кредитная и бюджетная политика Российской Федерации. В рамках ПС БР:

– по поручению ее участников (в том числе кредитных организаций, Федерального казначейства и его территориальных органов) осуществляются переводы денежных средств по счетам, открытым в Банке России;

– обеспечивается завершение расчетов по осуществляемым на территории Российской Федерации переводам денежных средств с использованием платежных карт;

– реализуется механизм завершения расчетов по сделкам, совершенным на финансовых рынках»[1].

В 2015 году через ПС БР было проведено 1398,5 млн платежных операций на сумму 1356,5 трлн рублей. Среднесуточный объем платежей в ПС БР в 2015 году составил 5,5 трлн рублей, среднесуточное количество таких платежей – 5,6 млн[2].

Создание современной модели прогнозирования вероятности дефолтов – трудоемкая задача, так как для работы с таким большим объемом транзакционных

---

[1] *Платежная система Банка России.*

[2] *Всемирный банк; Международная финансовая корпорация. 2016. Программа оценки финансового сектора Российской Федерации (Russian Federation Financial Sector Assessment Program): Пояснительная записка по финансовой инфраструктуре. Всемирный банк, Вашингтон, округ Колумбия. Всемирный банк.*

данных требуются значительные временные и вычислительные ресурсы. Поэтому на первом этапе мы не ставим перед собой задачу построения наилучшей модели. Цель данного исследования – проверить полезность транзакционных данных ПС БР для прогнозирования вероятности дефолта российских компаний.

Для достижения цели данного исследования на начальном этапе мы построили модели случайного леса для прогнозирования вероятности дефолта компании в каждой отраслевой группе с использованием стандартных годовых данных бухгалтерского учета. Далее мы добавили в модели признаки, созданные на основе транзакционных данных. Для проверки результатов мы также обучили модели логистической регрессии с L2-регуляризацией и взвешенными функциями правдоподобия. Кроме того, мы обучили модели случайного леса, используя только транзакционные данные, чтобы показать, что эти данные позволяют получать прогнозы вероятности дефолтов по крайней мере на три месяца раньше, чем в моделях на основе данных бухгалтерского учета.

Работа имеет следующую структуру. Глава 1 содержит краткий обзор литературы по прогнозированию вероятности дефолта. В главе 2 представлены источники данных, использовавшиеся в моделях. Модели машинного обучения, методы оптимизации гиперпараметров, примененные при поиске оптимальной архитектуры моделей, а также методика работы с несбалансированными выборками подробно рассмотрены в главе 3. Результаты моделей, рассчитанные для каждой отраслевой группы, и проверка их устойчивости представлены в главе 4. Заключение содержит основные результаты работы.

1. Обзор литературы

Работа Beaver (1966) стала ключевой в современной литературе по проблеме прогнозирования неплатежеспособности; впоследствии эта тема более подробно была исследована в работе Beaver (1968). Самая популярная модель в этом классе представлена в Altman (1968). Предложенные методы основывались на многомерной структуре с широким использованием многомерных моделей дискриминантного анализа (MDA) и построением стандартизированной оценки (Z-score) для разделения заемщиков на потенциально надежных и ненадежных. Однако критика нарушений статистических допущений, лежащих в основе подхода MDA, заставила исследователей в 1980-х годах сосредоточить свои усилия на разработке вероятностных моделей. Так, наиболее популярной является логит-модель, представленная в Ohlson (1980). Работа Odom and Sharida (1990) была одной из первых, в которой при прогнозировании банкротств использовалась нейронная сеть, состоящая из нескольких скрытых слоев. В качестве исходных данных были взяты финансовые коэффициенты, использованные в модели Альтмана (Altman).

Исследование Jackson and Wood (2013) показывает, что 25 различных методов, разработанных за последние 50 лет, дают разные результаты и по точности прогноза каждая из этих моделей превосходит более ранние.

В последнее десятилетие российские исследователи предпринимают усилия по построению моделей вероятности дефолта для российских компаний. Так, в работе Демешев и Тихонова (2014) сравниваются подходы к моделированию критического финансового положения средних и малых частных российских компаний в четырех отраслях (обрабатывающая промышленность, недвижимость, оптовая и розничная торговля, строительство) с использованием финансовых и нефинансовых показателей за 2011–2012 годы. Исследование основано на анализе годовых данных финансовой отчетности (баланс и отчет о прибылях и убытках) частных российских компаний в базе данных RUSLANA за 2011–2012 годы. В работе приведен перечень используемых финансовых и нефинансовых показателей. Из алгоритмов, выбранных авторами для сравнения (линейный дискриминантный анализ (LDA), квадратичный дискриминантный анализ, дискриминантный анализ смесей распределений, деревья классификации и случайный лес), наибольшую прогнозную силу показал алгоритм случайного леса. Авторы также отмечают, что среди нефинансовых показателей значимое влияние оказывают отрасль, федеральный округ и возраст предприятия. Размер предприятия менее важен, а его организационная форма практически не имеет значения. Среди финансовых показателей наиболее важными были показатели рентабельности, использования кредитных средств и ликвидности.

В статье Могилат (2015) исследуется банкротство компаний в России, идентифицированы и проанализированы основные тенденции и структурные характеристики компаний, признанных банкротами, и платежеспособных компаний за период 2007–2014 годов. Показано, что основными факторами прогнозирования вероятности дефолта являются чистая рентабельность активов, оборачиваемость общих активов, отношение чистой кредиторской задолженности компании к общим

активам и рентабельность активов в отрасли. Эти факторы остаются неизменно значимыми даже при изменении набора контрольных факторов и выборки. В следующей статье – Могилат (2019) – автор предлагает основанный на логистической регрессии эконометрический подход к оценке рисков финансовой устойчивости российских промышленных компаний, в котором учитывается как мировой опыт проведения таких исследований, так и особенности российских данных. Используются данные финансовой отчетности российских компаний за 2006–2016 годы. После процедуры фильтрации база данных содержит в среднем около 97 000 компаний в год. Для выявления банкротств анализируются данные за 2007–2017 годы. Факторы для моделирования финансовой устойчивости выбираются из мировой практики и на основе подхода, ранее предложенного в работе Могилат (2015). В исследовании также рассматривается проблема несбалансированных данных и способы ее решения. Для этого автор использует взвешенную функцию правдоподобия.

В другой статье – Karminsky and Burekhin (2019) – для прогнозирования вероятности дефолта российских компаний в строительной отрасли сравниваются такие алгоритмы, как логит- и пробит-модели, дерево решений, случайный лес и нейронные сети. Особое внимание уделяется особенностям построения моделей машинного обучения, влиянию дисбаланса данных на прогнозную силу моделей, анализу способов решения этих проблем, а также анализу влияния нефинансовых факторов. В настоящей работе используются показатели, основанные на публичной финансовой отчетности компаний за 2011–2017 годы. На этих данных рассчитываются финансовые показатели, отражающие экономическую деятельность компании. Кратко их можно описать как показатели прибыльности, ликвидности, деловой активности и финансовой стабильности.

Начиная с 1970-х годов основанные на бухгалтерской отчетности финансовые показатели являются важным источником данных для построения моделей вероятности дефолта. Однако информация, используемая для построения таких моделей, не учитывает волатильность результатов деятельности компании за анализируемый период, в связи с чем такой подход вызывает критику. Модель Мертона (Merton model), теоретические принципы которой являются основой для модели KMV, объясняет дефолт компании падением стоимости ее активов. В Merton (1974) предложена модель, в которой обыкновенные акции компании рассматриваются как опцион на ее активы. Основное преимущество опционного подхода заключается в том, что он позволяет оценить вероятность дефолта из ценовых значений, наблюдаемых на рынке. В то же время эта модель имеет существенный недостаток, особенно актуальный для России, – ограниченная база данных частных предприятий, акции которых находятся в обращении на фондовом рынке.

Для учета колебаний финансовых индексов компании в течение года и компаний, акции которых не торгуются на рынке, можно использовать данные платежных транзакций. На момент написания настоящей работы мы не обнаружили исследований по прогнозированию дефолта на данных о транзакциях российских компаний, хотя в литературе представлен ряд работ, которые рассматривают

связанные с этим вопросы. В статье Babaev et al. (2019) описывается использование моделей глубокого обучения на основе транзакционных данных для оценки кредитоспособности розничных клиентов. Авторы используют Embedding-Transactional RNN (ET-RNN, сетевая архитектура представлена в статье) на данных транзакции клиента (сумма платежа, тип карты, дата и время платежа, страна, валюта и тип перевода). Обучающая выборка состоит из транзакций примерно 740 000 клиентов. Общее количество транзакций составляет порядка 200 млн (около 800 транзакций на одного клиента). В качестве целевой переменной используется дефолт по потребительскому кредиту в течение года. В качестве базовых моделей – логистическая регрессия (с 400 факторами) и LightGBM (с 7000 генерируемыми факторами). В результате авторы показывают, что модель логистической регрессии имеет метрику ROC AUC 0,78, модель LightGBM – 0,81, а модель ET-RNN увеличивает этот показатель до 0,83. Следует отметить, что в отличие от классических методов, которые в значительной степени зависят от генерируемых факторов, метод ET-RNN не требовал ручной генерации факторов.

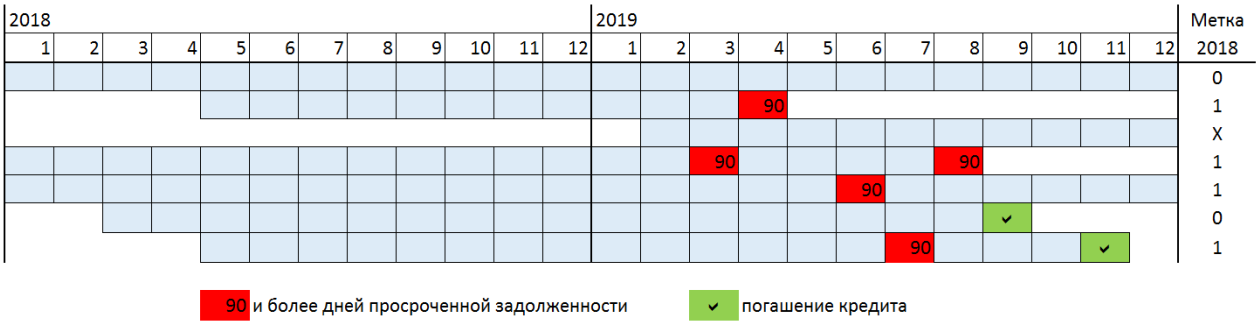
Возвращаясь к цели нашего исследования, отметим, что для проверки ценности данных платежной системы Банка России мы ориентируемся на модель случайного леса в качестве базовой и используем модель логистической регрессии для проверки устойчивости результатов. В качестве переменных для построения моделей вероятности дефолта мы рассматриваем стандартный набор финансовых показателей, которые более подробно описаны в следующей главе.

2. Данные

До того как прогнозировать вероятность дефолта компании, необходимо дать его определение. Согласно базельской методологии – Базель II (III) – на основе внутренних рейтингов (IRB) мы определяем дефолт как просрочку платежа более чем на 90 дней (BCBS 2017). Дата дефолта определяется с использованием информации о просроченной задолженности на основании данных кредитных бюро до 2018 года и формы отчетности 0409303 после 2018 года. Эта форма предоставляется ежемесячно, поэтому мы анализируем каждый период на предмет просроченной задолженности. Мы диагностируем дефолт, если продолжительность просрочки составляет 90 дней и более. Подробное определение дефолта представлено на рисунке 1 («X» в последнем столбце означает, что компания не входит в выборку).

Данные бухгалтерского учета за 2012–2018 годы взяты на сайте Росстата[3]. Поскольку наша цель – проверить полезность транзакционных данных ПС БР, которые доступны начиная с 2015 года, для расчета моделей мы используем данные бухгалтерского учета с 2015 года. Эти данные бухгалтерского учета были преобразованы в финансовые показатели для моделирования (Приложение, табл. 1).

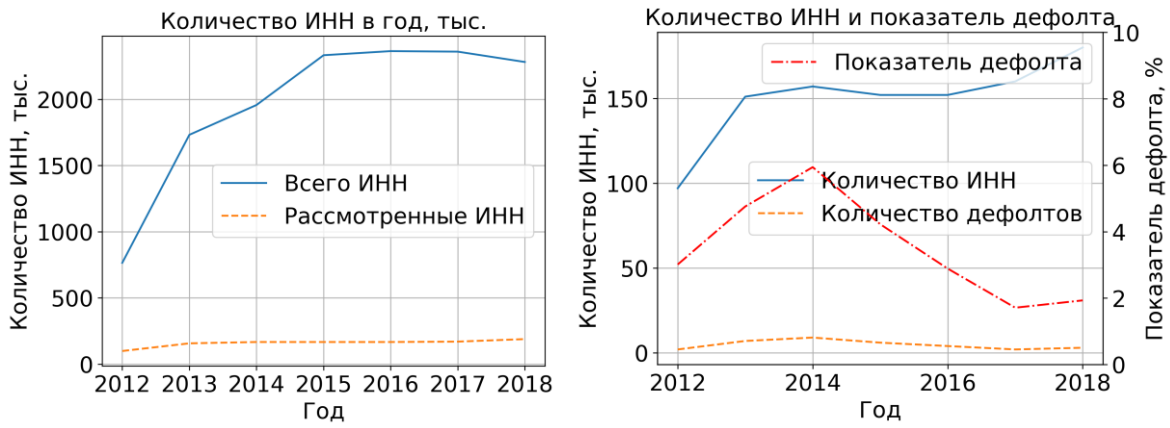
Рисунок 1. Определение дефолта компании



Первоначальная выборка данных варьировалась от 700 000 уникальных идентификационных номеров налогоплательщика (ИНН) в 2012 году до более 2 млн ИНН в 2018 году. Компании, которые имели кредит в течение года, следующего за отчетным периодом, были включены в набор данных (рассмотренные ИНН на рисунке 2). Показатель дефолта для всех включенных компаний варьировался в диапазоне 2–6% в год, что свидетельствует о дисбалансе классов в данных.

[3] Росстат.

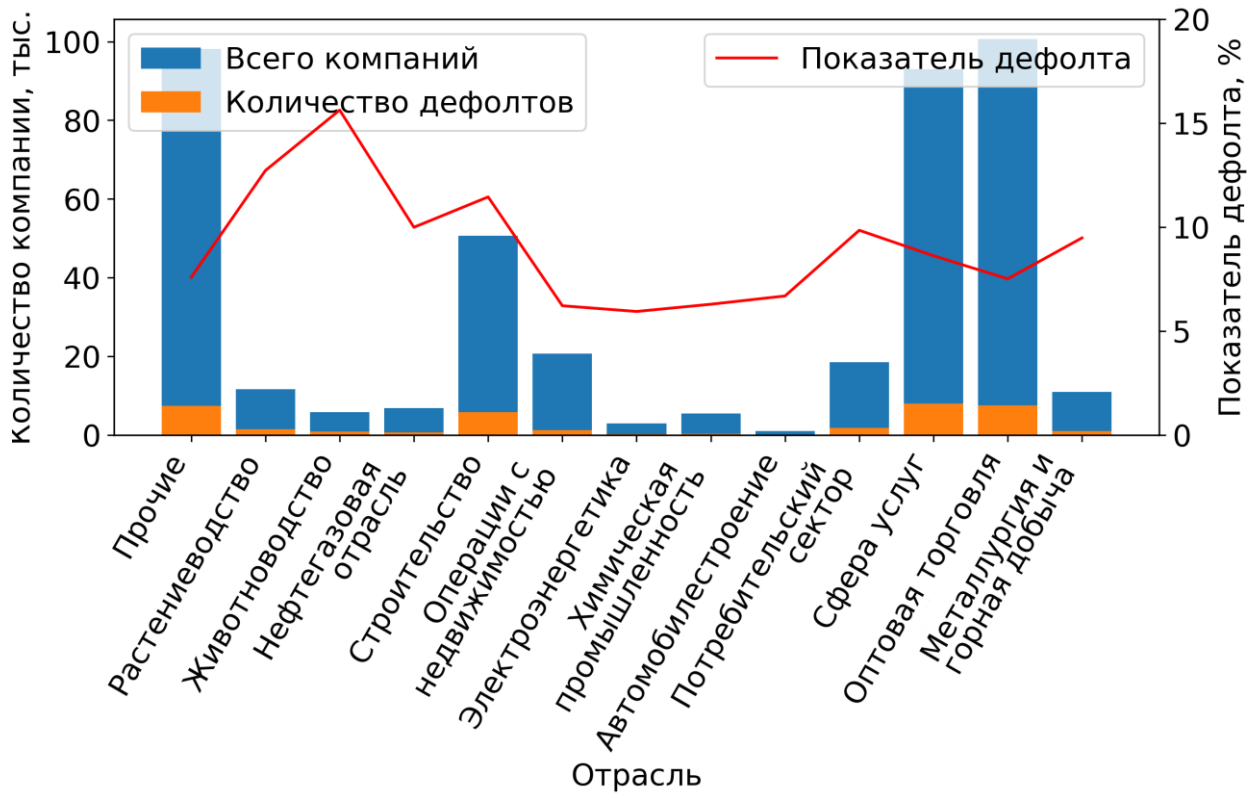
Рисунок 2. Количество ИНН в год и рассмотренных ИНН с показателем дефолта



Отраслевая принадлежность компаний определена по кодам ОКВЭД для основного вида деятельности. Основными критериями присвоения группы кодов ОКВЭД одной отраслевой группе являются идентичность набора факторов, влияющих на выручку и затраты компаний отрасли, и сходство степени и направления этого влияния.

На рисунке 3 показаны количество компаний и показатель дефолта в каждой отрасли. Существуют группы с небольшим количеством компаний, такие как электроэнергетика, химическая промышленность и автомобилестроение; эти группы усложняют прогнозирование вероятности дефолта. Возможны неточности в классификации по коду ОКВЭД, например некорректное определение головных офисов и филиалов компаний. Однако эта проблема не имеет серьезного значения для нашего исследования, поскольку мы не ставим цель построить наилучшую модель прогнозирования вероятности дефолта или учесть все возможные тонкости, поэтому этой проблемой можно смело пренебречь (подробное описание групп приведено в таблице 2 Приложения). Далее модели прогнозирования вероятности дефолта рассматриваются для этих отраслевых групп.

Рисунок 3. Количество компаний по отраслевым группам и показатель дефолта в 2012–2018 годах



Эта классификация позволяет определить показатель дефолта в разных отраслевых группах. Как правило, учет различий между отраслями помогает создавать более качественные модели. Наиболее проблемными отраслями по этому показателю в 2012–2018 годах были растениеводство и животноводство. Наименее подвержены риску дефолта оказались электроэнергетика и операции с недвижимым имуществом.

Данные платежной системы Банка России

Платежная система Банка России – системно значимая платежная система, через которую реализуется денежно-кредитная и бюджетная политика Российской Федерации. В этой системе денежные средства перечисляются через счета, открытые в Банке России (в том числе кредитных организаций, Федерального казначейства и его территориальных органов).

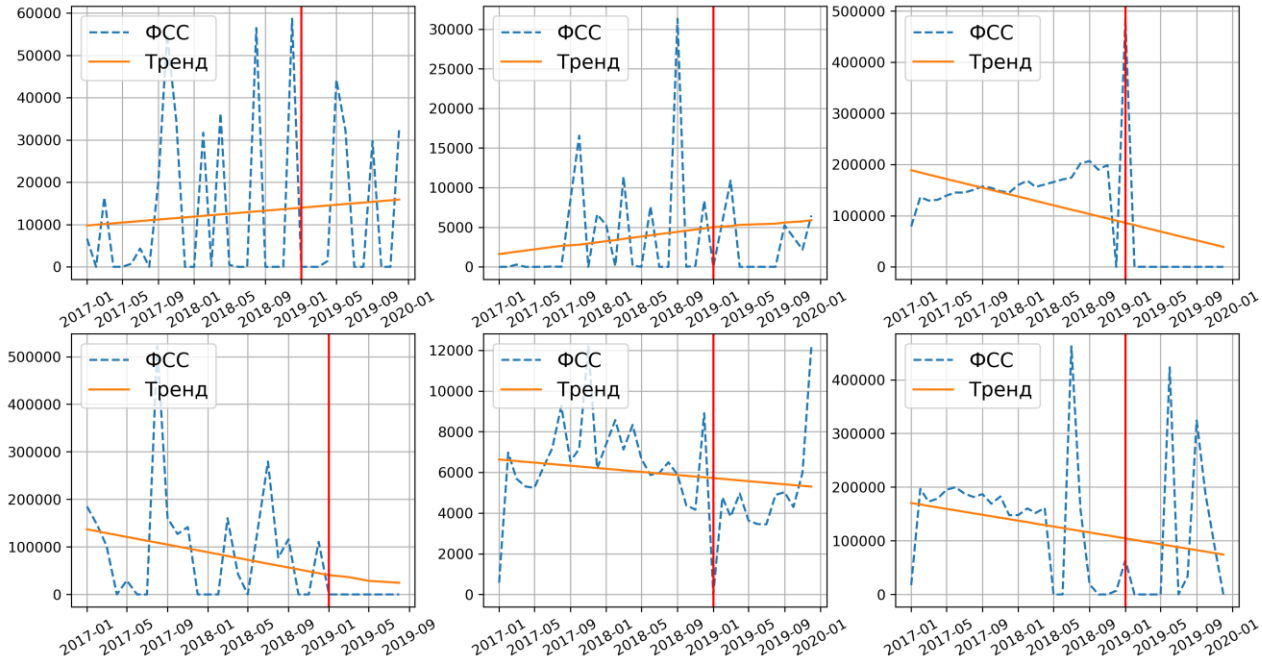
Данные об операциях по платежам могут использоваться в качестве дополнительной информации для оценки основных показателей деятельности фирмы[4]. Так, платежи по налогу на добавленную стоимость могут использоваться в качестве индикатора выручки компании, налог на прибыль – как доход организации,

[4] На транзакционные данные распространяется режим банковской тайны в соответствии с законодательством Российской Федерации о банках и банковской деятельности.

а налог на доходы физических лиц и страховые платежи могут быть индикатором заработной платы.

Основная цель настоящей работы – показать возможность использования и полезность этих данных для повышения качества стандартных моделей прогнозирования вероятности дефолта компаний. Для этого не требуется использовать всю имеющуюся информацию. Основная идея заключается в рассмотрении агрегированных данных (в основном налоговых платежей, но также рассматриваются и другие суммы, такие как общий приток и отток денежных средств; более подробное описание показателей приведено в таблице 3 Приложения). Мы используем данные ПС БР, доступные за 2015–2018 годы. На рисунке 4 приведен пример агрегированных ежемесячных данных по выплатам в Фонд социального страхования шести различных компаний с 2017 по 2019 год.

Рисунок 4. Пример данных ПС БР: агрегированные объемы ежемесячных платежей в Фонд социального страхования и их тренд для шести различных компаний с 2017 по 2018 год, отмеченных как допустившие дефолт в 2019 году



На первый взгляд может показаться, что если у компании возникают проблемы с возвратом кредитов, то она, скорее всего, сократит расходы – например, сокращая персонал и, соответственно, выплаты по социальному страхованию. Однако рисунок 4 показывает, что все не так просто. Только по поведению компании трудно определить, допустит она дефолт или нет. На рисунке отмечено, что каждая компания допустила дефолт в 2019 году, но все они ведут себя по-разному. Тем не менее в данных могут быть ценные линейные или нелинейные взаимосвязи, и для получения полезной информации мы будем использовать методы машинного обучения.

В базе данных ПС БР в 2019 году было зафиксировано около 10 млн транзакций в день, что действительно можно считать большими данными, и обработка такого большого объема может вызывать определенные сложности. Во избежание вычислительных трудностей мы агрегируем транзакционные данные по годам и кодам бюджетной классификации (для признаков, связанных с бюджетными платежами).

В качестве базовых признаков для прогнозирования вероятности дефолта мы используем стандартные характеристики, основанные на данных бухгалтерского учета, такие как отношение оборотного капитала к оборотным активам, отношение краткосрочной задолженности к выручке, коэффициент оборачиваемости дебиторской задолженности и другие (полный перечень 15 переменных приведен в таблице 1 Приложения). Расширенный набор признаков состоит из переменных, взятых из данных ПС БР, агрегированных по годам и нормализованных по активам (Приложение, табл. 3).

Все значения, отсутствующие в данных, заменяются нулями. Поскольку показатель может принимать чрезвычайно большие положительные или отрицательные значения для некоторых заемщиков (что существенно снижает качество моделей), значение каждого показателя ранжируется в рамках его отрасли. Например, отношение долга к прибыли от продаж принимает большие абсолютные значения, если сумма прибыли близка к нулю. Значения показателей затем нормализуются в соответствии с их отраслевой группой.

3. Методы

Наша задача в рамках настоящего исследования – рассмотреть метод прогнозирования вероятности дефолта, применить его к данным бухгалтерского учета, а затем добавить к этому методу переменные из ПС БР и сравнить результаты. Таким образом мы можем показать полезность данных ПС БР для прогнозирования вероятности дефолта.

Прогнозирование дефолтов компаний – задача бинарной классификации: допустит ли компания дефолт по своим кредитам в период, следующий за отчетным годом, или не допустит. То есть входные данные – это данные за отчетный год, а цель – определение отметки о дефолте в году, следующем за отчетным. Поэтому необходимо определить порог принятия решения о дефолте компании, который может привести к трудностям при сравнении моделей. Традиционно для этого используется площадь под ROC-кривой (ROC AUC), поскольку она свободна от такой проблемы.

В данной главе рассмотрены примененные методы машинного обучения, предсказывающие вероятность дефолта: модель логистической регрессии (раздел 3.1), модель случайного леса и ее метод определения важности признаков (раздел 3.2). В разделе 3.3 представлены детали реализации: схема кросс-валидации, методы оптимизации гиперпараметров и работы с несбалансированными данными.

3.1. Логистическая регрессия

Логистическая регрессия – наиболее часто используемый метод машинного обучения в моделях прогнозирования вероятности дефолта компаний. Одним из ключевых преимуществ алгоритма логистической регрессии является то, что модель может быть легко интерпретирована как функция входных данных. Модель состоит из коэффициентов для каждой переменной и константы, которые можно использовать для объяснения работы модели.

Мы используем логистическую регрессию с L2-регуляризацией. Чтобы найти параметры модели, нужно решить задачу минимизации:

$$\min_{w,c}\left(\frac{1}{2}w^T w + C\sum_{i=1}^{n}\log\left(\exp\left(-y_i(X_i^T w + c)\right) + 1\right)\right),$$

где $w = (w_0, \dots, w_p)$ – веса;

$X = (x_1, \dots, x_p)$ – входные данные;

C – обратная величина силы регуляризации;

y_i – целевая переменная.

Для решения проблемы минимизации мы используем алгоритм Limited-Memory BFGS (Broyden-Fletcher-Goldfarb-Shanno), описанный в Byrd et al (1995) и реализованный в библиотеке scikit-learn для Python[5].

[5] *Реализация модели логистической регрессии в библиотеке scikit-learn для Python.*

3.2. Случайный лес

Случайный лес – метод машинного обучения, заключающийся в использовании ансамбля решающих деревьев, предложенный Лео Брейманом в Breiman (2001). Каждое дерево представляет собой модель древовидной структуры с узлами в качестве точек принятия решения, которые задают правила для объясняющей переменной, чтобы прогнозировать целевую переменную. Разделение узлов дерева принятия решений основано на конкретном критерии для одной из переменных признака, например Gini (для классификации) или суммы квадратов (для регрессии), из всего набора данных. Листовой узел, также называемый терминальным, содержит подмножество наблюдений. Разделение продолжается до тех пор, пока не будет сформирован листовой узел.

Преимущество метода случайного леса заключается в том, что с его помощью можно обрабатывать большие наборы данных с более высокой размерностью и что он имеет высокую точность. Этот метод сложнее интерпретировать по сравнению с логистической регрессионной моделью.

В нашем исследовании мы используем реализацию данного алгоритма из библиотеки scikit-learn для Python[6]. Гиперпараметры для оптимизации были выбраны из числа общепринятых на основе работы Luppe (2014).

Случайные леса могут использоваться для ранжирования важности признаков в задачах регрессии или классификации. Методика описана в работе (Breiman) (2001) и реализована в пакете scikit-learn для Python. Этот метод широко используется в том числе для экономических задач, таких как рассмотренные в работе Chakraborty and Joseph (2017)[7].

В процессе обучения ошибки вне выборки для каждого результата обработки данных усредняются по всему лесу. Чтобы определить важность *i*-го признака, сначала мы обучаем модель, затем переставляем значения *i*-го признака среди обучающих данных, а затем повторно оцениваем ошибку вне выборки. Важность *i*-го признака рассчитывается путем усреднения разницы в ошибке вне выборки до и после перестановок по всем деревьям.

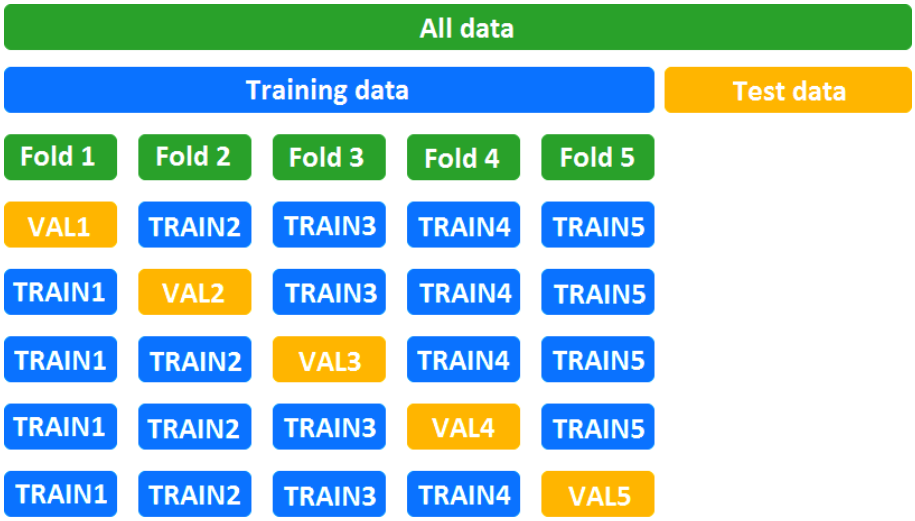
3.3. Детали реализации

Для получения надежных результатов и решения проблемы переобучения мы использовали *k*-блочную кросс-валидацию с сохранением соотношения классов в подмножествах. Для тестирования результатов модели мы выделяем 1/5 полного набора данных (рис. 5). Набор данных обучения разделен на пять подмножеств, каждое из которых используется в качестве контрольной выборки. Остальные тренировочные данные применяются для обучения модели с заданным набором гиперпараметров. Для каждого подтеста мы оцениваем метрику ROC AUC и

[6] *Реализация классификатора случайный лес (Random Forest Classifier) в библиотеке scikit-learn для Python.*
[7] *Однако, как отмечено в Strobl et al (2007), важность признака на основе данного алгоритма может ввести в заблуждение в случае с признаками с большим количеством уникальных значений.*

выбираем лучшую модель с заданным набором гиперпараметров. Для оценки модели на тестовых данных мы используем всю тренировочную выборку (4/5 от полного набора данных) с выбранными гиперпараметрами.

Рисунок 5. Стандартная схема k-блочной кросс-валидации



Все сравнения результатов моделей, приведенные в данной работе, проводятся на тестовом наборе в соответствии с представленной схемой кросс-валидации.

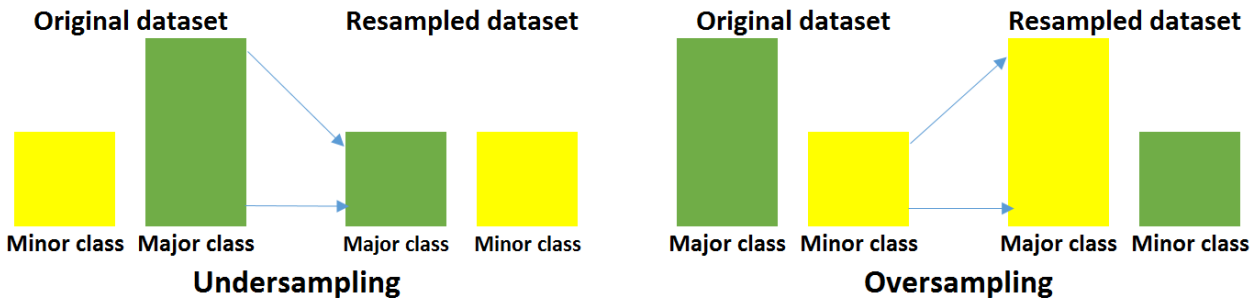
Настройка гиперпараметров – важная часть процесса машинного обучения. Поиск по сетке – один из наиболее часто используемых методов оптимизации гиперпараметров, но итерации с большим количеством параметров занимают много времени. В работе Bergstra and Bengio (2012) эмпирически и теоретически показано, что метод со случайным выбором гиперпараметров более эффективен, чем перебор по сетке.

Исследователи в Dewancker et al. (2016) предлагают метод байесовской оптимизации. Байесовская оптимизация – алгоритм последовательной оптимизации, в котором для повышения производительности модели используются результаты предыдущей итерации для определения последующих значений гиперпараметров. При таком подходе сокращается количество точек и время, необходимое для поиска наилучших гиперпараметров. Мы используем этот метод, реализованный в библиотеке scikit-learn для Python[8].

Как описано в главе 2, мы работаем с несбалансированным набором данных. Многие алгоритмы машинного обучения чувствительны к распределению классов в наборе данных. Поэтому такие модели имеют низкое качество прогнозирования. Наиболее популярными решениями проблемы несбалансированной классификации являются методы, изменяющие распределение классов (Ganganwar (2012); He and Ma (2013); Sonak and Patankar (2015) путем случайного удаления (undersampling) примеров из превалирующего класса или путем дублирования (oversampling) примеров класса с наименьшим представительством в обучающей выборке (рис. 6).

[8] *Реализация байесовского оптимизатора (Bayesian optimiser) в библиотеке scikit-learn для Python.*

Рисунок 6. Пример методов балансировки обучающей выборки



При применении метода с удалением части выборки в нашей задаче значительно уменьшается размер обучающей выборки и увеличивается ошибка прогнозирования. Метод дублирования более надежен для нашей задачи, но он значительно увеличивает набор данных за счет повторения элементов и увеличивает вычислительное время на обучение моделей. В нашем случае хорошо подходит метод взвешенной функции правдоподобия для логистической регрессии, описанный в King and Zeng (2001). Этот алгоритм реализован в библиотеке scikit-learn[9]. В моделях случайного леса для настройки весов мы используем обратно пропорциональную частоту попадания в группу во входных данных (подробнее см. работу Chen et al., 2003). Этот алгоритм реализован в библиотеке scikit-learn[10].

[9] *Реализация модели логистической регрессии в библиотеке scikit-learn для Python.*
[10] *Реализация модели классификатора случайного леса (Random Forest Classifier model) в библиотеке scikit-learn для Python.*

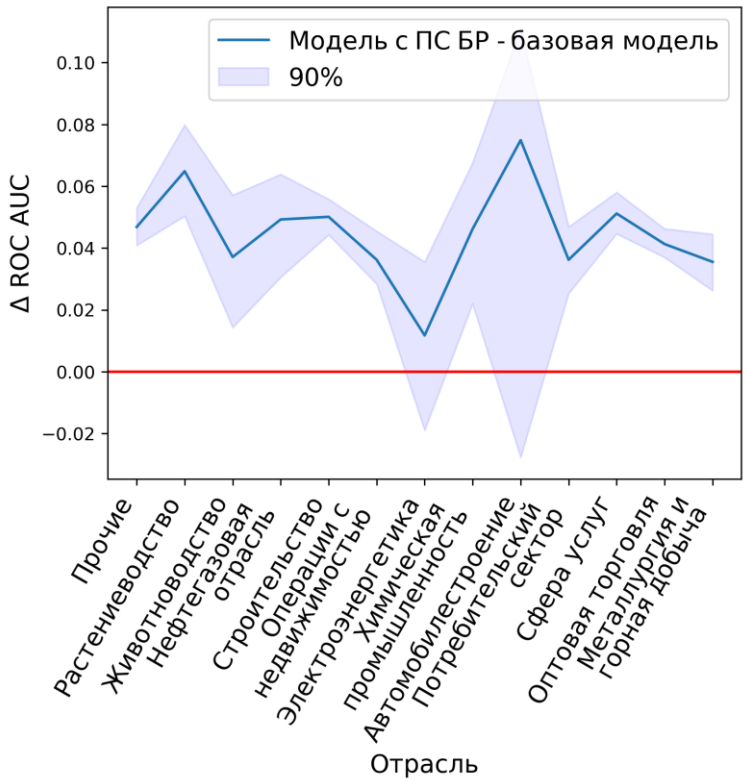
4. Результаты

В настоящем разделе обобщаются результаты анализа модели. В разделе 4.1 мы рассматриваем разницу в качестве по метрике ROC AUC для прогнозов вероятности дефолта с использованием моделей случайного леса с данными бухгалтерского учета (базовая модель) и моделей случайного леса с признаками, расширенными при помощи ПС БР для каждой отраслевой группы. В разделе 4.2 представлены результаты сравнения моделей прогнозирования дефолта с использованием данных только ПС БР и базовой модели. Оценки важности признаков, полученных с помощью моделей случайного леса, представлены в разделе 4.3. Раздел 4.4 содержит результаты проверки устойчивости результатов с помощью моделей логистической регрессии с использованием L2-регуляризации и взвешенной функцией правдоподобия. Все результаты показаны для тестового набора данных согласно схеме *k*-блочной кросс-валидации.

4.1. Модель случайного леса

Чтобы проверить полезность данных ПС БР, мы построили модели случайного леса с использованием данных бухгалтерского учета для прогнозирования вероятности дефолта компаний по каждой отраслевой группе. Затем мы построили модели вероятности дефолта для каждой отраслевой группы, используя одни и те же инструменты, добавив данные ПС БР. При оптимизации гиперпараметров каждой модели мы использовали метод *k*-блочной кросс-валидации. В результате показатели метрики ROC AUC составили в среднем 75 и 79% по отраслевым группам для базовой и расширенной моделей соответственно. На рисунке 7 показана разница метрик ROC AUC на тестовом наборе данных для каждой отраслевой группы с 90%-ным доверительным интервалом, полученным с использованием бутстрэпа (bootstrap). Результаты в абсолютных значениях метрики ROC AUC показаны на рисунке 1 в Приложении.

Рисунок 7. Разница метрики ROC AUC на тестовом наборе данных между моделями случайного леса с использованием расширенных данных ПС БР и моделями случайного леса с использованием бухгалтерских данных

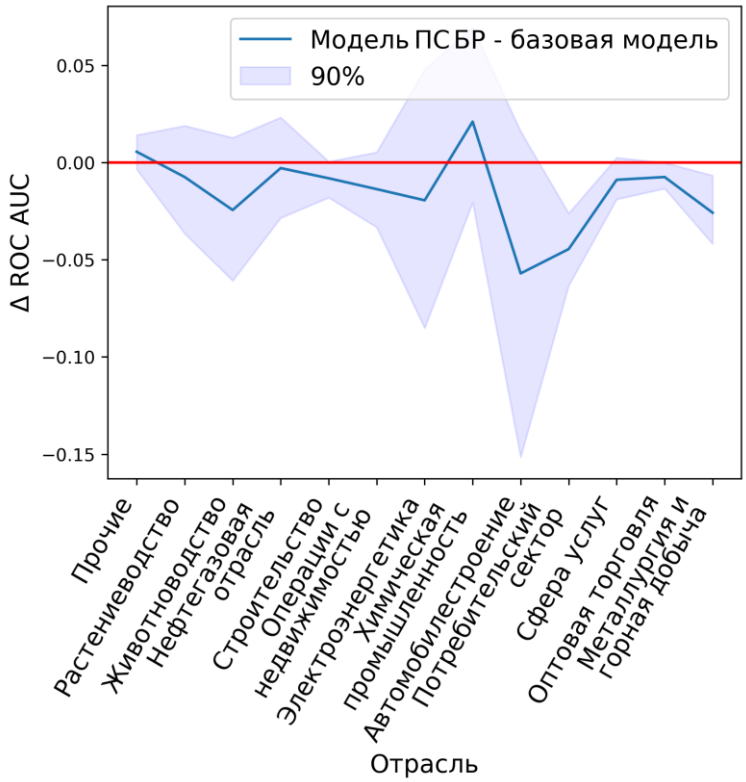


На графике показано, что метрика ROC AUC для прогнозов моделей с расширенным набором переменных в среднем на 4 пункта выше базовых моделей по всем группам отраслей и для всех групп эта разница статистически значима (рис. 7). Это свидетельствует о том, что данные ПС БР полезны и информативны и могут быть использованы для повышения качества моделей прогнозирования вероятности дефолта российских компаний.

4.2. Модель случайного леса на основе данных ПС БР

Кроме того, мы обучили модели случайного леса для прогнозирования вероятности дефолта компаний в каждой отраслевой группе с использованием только данных ПС БР. В результате мы получили показатели метрики ROC AUC 75 и 74% в среднем по отраслевым группам для базовой и ПС БР моделей соответственно. На рисунке 8 показана разница метрики ROC AUC на тестовом наборе между моделями, обученными только на транзакционных данных, и моделями, обученными на данных бухучета для каждой отраслевой группы.

Рисунок 8. Разница метрики ROC AUC на тестовом наборе данных между моделями случайного леса с использованием бухгалтерских данных и моделями случайного леса с использованием только данных ПС БР

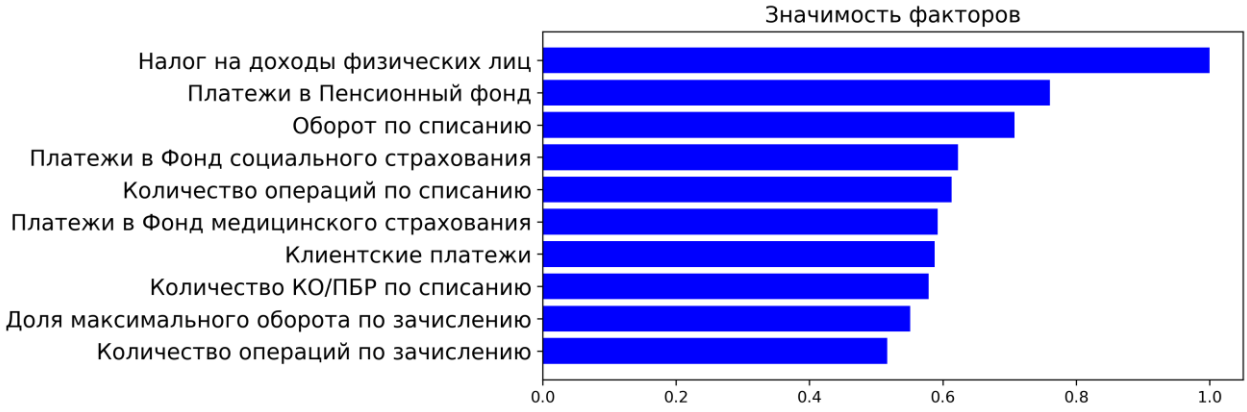


На графике показано, что метрика ROC AUC для прогнозов модели на основе данных платежной системы в среднем на 1 пункт ниже. Тем не менее эти прогнозы можно получить как минимум на три месяца раньше, учитывая задержку в публикации годовой бухгалтерской отчетности. Это дает больше времени для принятия решений и больше шансов выйти из ситуации, в которой компания имеет высокую вероятность дефолта.

4.3. Важность факторов методом случайного леса

Для оценки важности факторов ПС БР мы использовали модель случайного леса, обученную только на данных ПС БР. На рисунке 9 показаны 10 наиболее значимых факторов (с нормализацией по максимальной значимости)[11].

Рисунок 9. Топ-10 наиболее значимых факторов по оценкам случайного леса

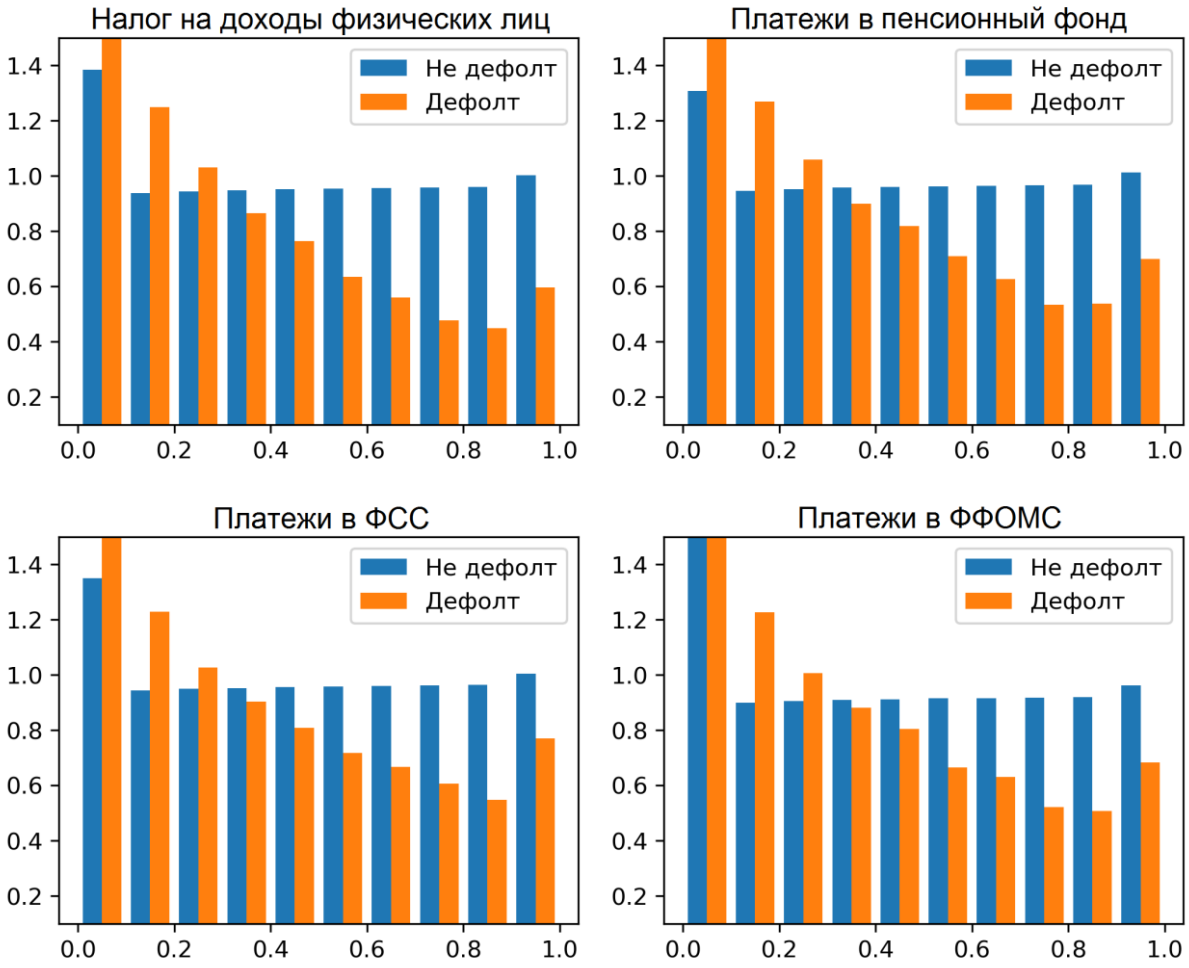


На основании таких оценок можно сказать, что факторы, относящиеся к заработной плате (платежи по НДФЛ, пенсионные отчисления и отчисления в фонд социального страхования), сумма и количество исходящих платежей имеют большое значение для прогнозирования вероятности дефолта компаний. Этот вывод подтверждает, что компании сокращают свои расходы на персонал, если у них возникают проблемы с выплатами по кредитам, хотя так поступает не каждая компания (рис. 4).

На рисунке 4 не отражены четкие закономерности в динамике выплат заработной платы фирмами, у которых возникли проблемы с возвратом кредитов, но результаты, представленные в данном разделе, показывают, что это наиболее важные признаки. Чтобы увидеть различия в данных показателях, мы построили гистограммы плотности их распределения (рис. 10).

[11] *Все учетные переменные входят в топ-30 по алгоритму важности признака «случайный лес» (тренировка на данных бухучета и данных ПС БР).*

Рисунок 10. Гистограммы плотности распределения нормированных показателей налога на доходы физических лиц, выплат в пенсионный фонд, выплат в Фонд социального страхования и выплат в Фонд медицинского страхования



Предоставленные гистограммы показывают, что распределение этих признаков имеет различные формы для компаний, которые отмечены как допустившие и не допустившие дефолт в следующему году. Эта разница – именно та информация, которая извлекается из данных при помощи наших моделей.

4.4. Проверка устойчивости с использованием модели логистической регрессии

Для проверки устойчивости и подтверждения результатов мы также построили модели вероятности дефолта компаний для каждой отраслевой группы, используя метод логистической регрессии с L2-регуляризации и взвешенной функцией правдоподобия, сначала на основе данных бухучета, а затем с использованием расширенного набора признаков из данных ПС БР. Поскольку расширенная модель не смогла успешно обработать такое большое количество переменных даже при L2-регуляризации, мы предварительно отобрали генерируемые признаки из данных

ПС БР с использованием метода важности признаков случайного леса. Кроме того, мы оптимизировали гиперпараметр обратной силы регуляризации для каждой модели, используя *k*-блочную кросс-валидацию. В результате мы получили показатели метрики ROC AUC 74 и 76% в среднем по отраслевым группам для базовой и расширенной моделей соответственно. На рисунке 11 показана разница в метрике ROC AUC для каждой группы с 90%-ным доверительным интервалом, полученным с использованием бутстрэпа.

Рисунок 11. Разница метрики ROC AUC на тестовом наборе данных между моделями логистической регрессии с использованием расширенных признаков ПС БР и моделями логистической регрессии использованием бухгалтерских данных

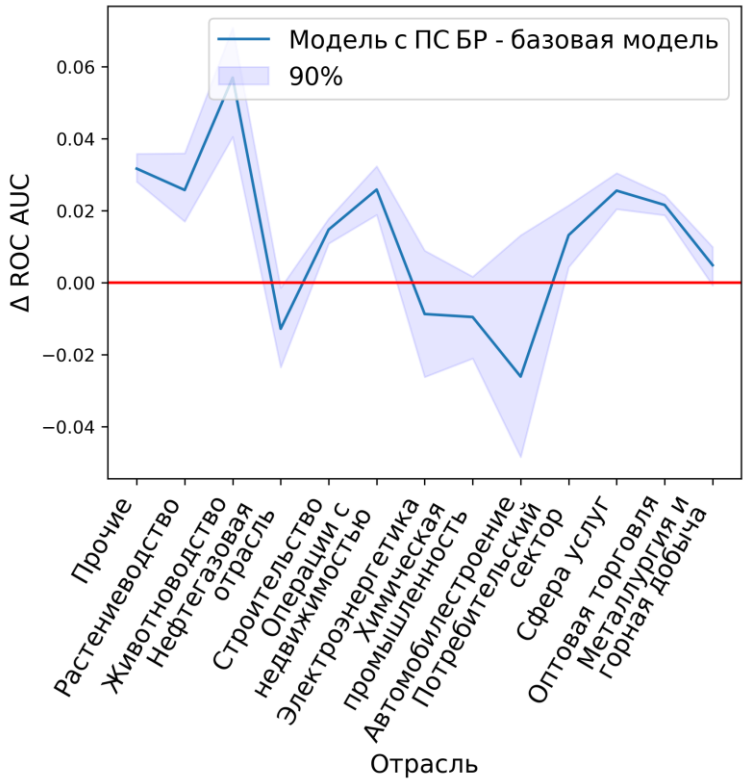


График показывает, что метрика ROC AUC для моделей с расширенным набором признаков почти везде выше, чем в базовых моделях, за исключением нефтяной, газовой, электроэнергетической, химической и автомобильной промышленности, где объем данных в моделях невелик и результаты очень волатильны. Согласно метрике ROC AUC, прогнозное качество расширенных моделей в среднем на 2 пункта выше по всем отраслевым группам. Результаты показывают важность данных ПС БР, как и в моделях случайного леса. Разницу в результатах моделей случайного леса и логистической регрессии можно объяснить небольшим количеством компаний в некоторых группах (и, как следствие, более значительной разницей), а также более низкой чувствительностью случайного леса к отклонениям в наборе данных.

Заключение

Основная цель этой работы – продемонстрировать полезность информации из ПС БР для прогнозирования вероятности дефолта российских компаний. Для достижения этой цели мы оценили модели с использованием методов машинного обучения и показали, что данные ПС БР, добавленные к данным бухгалтерского учета, улучшают качество прогнозирования согласно метрике ROC AUC.

Таким образом, данные ПС БР содержат дополнительную полезную информацию. Метод важности признаков случайного леса показывает, что основными источниками дополнительной информации являются налоги на заработную плату и социальные выплаты.

Кроме того, может быть полезным использование ПС БР в качестве самостоятельного источника данных. Несмотря на то, что результаты наших моделей, обученных только на данных ПС БР, по метрике ROC AUC несколько уступают моделям, основанным на данных бухгалтерского учета, данные ПС БР создают возможность получать оценки вероятности дефолта раньше, что может быть особенно важно в современном быстро меняющемся мире.

Одним из направлений будущих исследований может быть анализ данных платежных операций при помощи методов, основанных на нейронных сетях, а также оценка вероятности дефолта компаний на ежемесячной основе.

Список литературы

1. Altman E.I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609.

2. Babaev D., Savchenko M., Tuzhilin A. and Umerenkov D. 2019, July. E.T.-RNN: Applying Deep Learning to Credit Loan Applications. In Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19). 8 pages.

3. Bergstra J. and Bengio Y. 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(2), pp. 281–305.

4. Beaver W. 1966. Financial ratios as predictors of failure. Journal of Accounting Research, 4(3) (Supplement), pp. 71–111.

5. Beaver W. 1968. Market prices, financial ratios and the prediction of failure. Journal of Accounting Research, 6(2), pp. 179–192.

6. Breiman L. 2001. Random forests. Machine learning 45(1), pp. 5–32.

7. Byrd R.H., Lu P., Nocedal J. and Zhu C. 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific and Statistical Computing, 16(5), pp. 1190–1208.

8. Chakraborty C. and Joseph A. 2017. Machine learning at central banks. Bank of England. Working Paper № 674. 89 pages.

9. Chen C., Liaw A. and Breiman L. 2003. Using random forest to learn imbalanced data. University of California, Berkeley, 110(1–12), p. 24.

10. Dewancker I., McCourt M. and Clark S. 2016. Bayesian Optimization for Machine Learning: A Practical Guidebook.

11. Ganganwar V. 2012. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4), pp. 42–47.

12. He H. and Ma Y. Eds. 2013. Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons. Technology & Engineering. 216 pages.

13. Hastie T., Tibshirani R. and Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer. 745 pages.

14. Jackson R. and Wood A. 2013. The performance of insolvency prediction and credit risk models in the UK: A comparative study. The British Accounting Review, 45(3), pp. 183–202.

15. Karminsky A.M. and Burekhin R.N. 2019. Comparative analysis of methods for forecasting bankruptcies of Russian construction companies. *Business Informatics, 13*(3), pp. 52–66.

16. King G. and Zeng L. 2001. Logistic regression in rare events data. *Political Analysis, 9*(2), pp. 137–163.

17. Louppe G. (2014) Understanding Random Forests: From Theory to Practice. PhD Thesis, University of Liege.

18. Merton R.C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *T*he Journal of Finance, 29(2), pp. 449–470.

19. Odom M.D. and Sharda R. 1990. A neural network model for bankruptcy prediction. In 1990 IJCNN International Joint Conference on neural networks, pp. 163–168. IEEE.

20. Ohlson J.A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, pp. 109–131.

21. Shibitov D. and Mamedli M. 2019. The finer points of model comparison in machine learning: forecasting based on Russian banks' data. Bank of Russia Working Paper Series, 43.

22. Sonak A. and Patankar R.A. 2015. A survey on methods to handle imbalance dataset. Int. J. Comput. Sci. Mob. Comput., 4(11), pp. 338–343.

23. Strobl C., Boulesteix A.L., Zeileis A. and Hothorn T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(1), pp. 1–21.

24. Демешев Б.Б., Тихонова А.С. 2014. Прогнозирование банкротства российских компаний: межотраслевое сравнение // Экономический журнал Высшей школы экономики, 18(3), с. 359–386.

25. Могилат А.Н. 2015. Банкротство компаний реального сектора в России: основные тенденции и финансовый «портрет» типичного банкрота // Научные труды ИНП РАН, 13, с. 156–186.

26. Могилат А.Н. 2019. Моделирование финансовой устойчивости компаний реального сектора (на примере промышленности) // Вопросы экономики (3), с. 101–118.

Приложение

Таблица 1. Признаки базовой модели

Название	Расшифровка
K1	Коэффициент текущей ликвидности
K2	Рентабельность активов
K3	Валовая рентабельность
K4	Рентабельность по чистой прибыли
K5	Операционная рентабельность
K6	Коэффициент автономии
K7	Задолженность/прибыль от продаж
K8	Коэффициент покрытия процентов
K9	Коэффициент процентной нагрузки
K10	Заемные средства/капитал
K11	Соотношение оборотного капитала и оборотных активов
K12	Коэффициент налоговой нагрузки
K13	Краткосрочный долг/выручка
K14	Коэффициент оборота дебиторской задолженности
K15	Коэффициент оборота кредиторской задолженности

Таблица 2. Перечень кодов ОКВЭД по отраслям[12]

Наименование отрасли	Коды ОКВЭД	Задолженность[13] на 01.01.2019, млрд руб.
АПК – растениеводство	01, 1	670
АПК – животноводство	01, 4	1 009
Нефтегазовая промышленность	06.1, 06.2, 19.2, 46.71, 46.12.1, 49.50.1, 49.50.2	3 253
Строительство	22.23, 25.11, 41, 42, 43	1 501
Операции с недвижимостью	68	2 764
Электроэнергетика и ЖКХ	35	1 302
Химическая промышленность	20, 21, 22	1 081
Автомобилестроение	29	293
Потребительский сектор	10, 11, 12, 13, 14, 18, 31, 32	1 165
Сфера услуг	45, 47, 49, 50, 51, 52, 53, 58, 61, 62, 63	3 888
Оптовая торговля (за исключением топлива и полезных ископаемых)	46	2 128
Металлургия и горная промышленность	05, 07, 08, 09, 23, 24, 46.72, 46.12.2	2 910
Прочие	Прочие	5 921
Всего		**27 885**

[12] Заемщики, которым был присвоен ОКВЭД 64.20 «Деятельность холдинговых компаний», были отнесены экспертами к одной из отраслей, указанных в таблице 2.
[13] По данным формы отчетности 0409303.

Таблица 3. Данные платежной системы Банка России

№	Наименование	Расшифровка
1	DT	Дата отчета
2	INN	Идентификационный номер налогоплательщика
3	OKVED_CODE	ОКВЭД
4	CNT_FI_ID_DB	Количество КО/ПБР[14] по списанию
5	CNT_FI_ID_CR	Количество КО/ПБР по зачислению
6	PRC_MAX_DB	Доля максимального оборота по списанию в КО/ПБР
7	PRC_MAX_CR	Доля максимального оборота по зачислению в КО/ПБР
8	TOT_DB	Оборот по списанию
9	TOT_CR	Оборот по зачислению
10	CNT_DB	Количество операций по списанию
11	CNT_CR	Количество операций по зачислению
12	DB06	Земельный налог
13	DB07	Налог на игорный бизнес
14	DB08	Налог на имущество
15	DB09	Транспортный налог
16	DB10	Прочие налоги
17	DB11	Налог на прибыль организаций
18	DB12	Налог на доходы физических лиц
19	DB13	НДС на товары, ввозимые на территорию России
20	DB14	НДС на товары (работы, услуги), реализуемые на территории России
21	DB15	Налог, взимаемый в связи с применением упрощенной системы налогообложения
22	DB16	Единый налог на доход для отдельных видов деятельности
23	DB17	Единый сельскохозяйственный налог
24	DB18	Налог, взимаемый в связи с применением патентной системы налогообложения
25	DB20	Прочие платежи на социальные нужды
26	DB21	Платежи в Пенсионный фонд Российской Федерации
27	DB22	Платежи в ФСС Российской Федерации
28	DB23	Платежи в ФФОМС
29	DB24	Платежи в пользу ФТС
30	DB30	Прочие платежи в бюджет
31	DB31	Клиентские платежи
32	DB32	Платежи нерезидентам
33	DB33	Расчеты с биржей
34	DB35	Списания с депозитов
35	DB36	Платежи в пользу банков нерезидентов
36	DB37	Покупка иностранной валюты
37	DB38	Погашение кредита
38	DB00	Прочие платежи по списанию
39	CR39	Возврат НДС
40	CR40	Платежи из бюджета
41	CR41	Клиентские платежи
42	CR42	Платежи от нерезидентов
43	CR43	Расчеты с биржей
44	CR45	Зачисления на депозиты

[14] КО/ПБР – кредитная организация / подразделение Банка России.

№	Наименование	Расшифровка
45	CR46	Платежи от банков нерезидентов
46	CR47	Продажа иностранной валюты
47	CR48	Получение кредита
48	CR00	Прочие платежи по зачислению

Рисунок 1. Метрика ROC AUC моделей на тестовом наборе данных с разными обучающими признаками