Предсказание успешности выполнения контрактов по государственным закупкам

М. А. ХРОМОВ
Московский государственный университет им. М. В. Ломоносова
e-mail: m@khromov.su

А. В. ШОКУРОВ
Московский государственный университет им. М. В. Ломоносова
e-mail: shokurov.anton.v@yandex.ru

УДК 004.89

Ключевые слова: успешность выполнения контракта, государственные закупки, поиск важных признаков.

Аннотация

В работе исследуется задача предсказания успешности выполнения контрактов, заключающихся в России. За основу взят алгоритм машинного обучения — градиентный бустинг над решающими деревьями. Производится настройка параметров классификатора, генерация и поиск наиболее значимых признаков. Были найдены следующие важные признаки: процент падения цены контракта; стоимость контракта в день; цена контракта на одного работника; цена контракта, умноженная на изменение цены.

Abstract

M. K. Khromov, A. V. Shokurov, Determining fulfilment rate of public procurement contracts, Fundamentalnaya i prikladnaya matematika, vol. 25 (2024), no. 1, pp. 237—250.

This paper investigates the problem of predicting the success of contracts concluded in Russia. It is based on a machine learning algorithm: a gradient binning over solver trees. The classifier parameters are adjusted, and the most important features are generated and searched for. The following important attributes were found: percentage of contract price drop; contract price per day; contract price per employee; contract price multiplied by price change.

1. Введение

В работе рассматривается проблема определения вероятности успешности выполнения контракта между заказчиком и поставщиком с применением традиционных методов машинного обучения [10]. Целью работы является определение важных признаков и параметров классификатора, определяющего вероятность успешности контракта. Мы будем рассматривать контракты на предоставление товаров или услуг между юридическими лицами. Каждая из сторон

Фундаментальная и прикладная математика, 2024, том 25, № 1, с. 237—250.

может быть как государственной компанией, так и коммерческой. Процесс оценки вероятности успешного выполнения контракта называется скоринг [7]. Тема скоринга контрактов весьма важная и актуальная. Она полезна банкам или иным кредитным организациям, выдающим тендерные кредиты и банковские гарантии.

Тендерный кредит — это специальный кредит, получаемый участником тендера/конкурса/электронного аукциона, т. е. участником размещения государственного и муниципального заказа в соответствии с требованиями Федерального закона от 05.04.2013 № 44-ФЗ (редакция от 01.05.2019) «О контрактной системе в сфере закупок товаров, работ, услуг для обеспечения государственных и муниципальных нужд». В случае если организация решила участвовать в тендере, но на её счетах недостаточно средств для обеспечения заявки или организация не хочет отвлекать значительные средства из оборота, организация может обратиться за получением тендерного кредита.

Банковская гарантия — один из способов обеспечения исполнения обязательств, при котором банк, иное кредитное учреждение, страховая организация или иная коммерческая организация (гарант) выдаёт по просьбе должника (принципала) письменное обязательство уплатить кредитору (бенефициару) денежную сумму при предоставлении им требования об её уплате. Если гарантию выдаёт банк или иная кредитная организация, то она именуется банковской гарантией.

У контракта может быть несколько исходов:

— исполнен (контракт был выполнен в полном объёме);
— частично исполнен (могут быть наложены штрафы и пени);
— расторгнут (по решению суда или по соглашению сторон).

Неуспешно выполненным контрактом будем считать контракт, который был расторгнут по решению суда. Остальные будем считать успешными, даже если были наложены штрафы.

При рассмотрении банком заявки на получение тендерного кредита или банковской гарантии на выполнение того или иного контракта ему очень важно знать, окажется контракт дефолтным или же он будет успешно выполнен. Стоимость банковской гарантии — в среднем около 1 % от суммы контракта. Если же контракт расторгается, то по условиям банковской гарантии банк обязан выплатить заказчику 30 % от суммы контракта.

В России на электронных торговых площадках размещается несколько миллионов контрактов в год, поэтому данных для обучения систем машинного обучения достаточно много, хоть их и непросто получить.

2. Машинное обучение

Так же как человек учится чему-то на примерах, методы машинного обучения способны «обучать» программу решать поставленную задачу, но для этого

им нужно подготовить большой массив данных. Таким образом, в случаях, где сложно составить определённый алгоритм решения, полезно использовать машинное обучение. В частности, оно хорошо справляется и с поставленной в этой работе задачей.

Традиционно скоринг проходит по определённому алгоритму. Просматриваются определённые параметры контракта и отбрасываются те контракты, которые не подпадают под критерии, заданные банком. Критерии выбираются вручную специальными работниками банка, которые регулярно обновляют их, основываясь на последнем опыте выдачи гарантий. Это довольно неточный способ проверки, так как критериев очень много и результат исполнения контракта может зависеть сразу от многих из них в совокупности и нелинейно. Методы машинного обучения же могут находить сложные зависимости при правильной настройке параметров и достаточно большом количестве данных для обучения.

На протяжении всей работы был использован один вид классификатора — классификатор, использующий метод градиентного бустинга [8] над решающими деревьями. Это один из наиболее популярных и эффективных методов классификации. В этом исследовании была проведена работа по сбору данных, их обработке, выявлению наиболее важных признаков и настройке параметров классификатора.

3. Данные

Сбор данных для решаемой задачи был одной из самых больших проблем в проведённом исследовании из-за сложности их получения. На данный момент в России работает государственный портал, на котором собирается информация по контрактам — `zakupki.gov.ru`. Была проведена попытка собрать данные с него, но автоматический сбор был заблокирован со стороны портала, а сотрудники отказались предоставлять возможность получить данные по контрактам. Но электронная торговая площадка «Сбербанк АСТ» [5] предоставила информацию по имеющимся у неё контрактам, будучи заинтересованной в такой программе скоринга. Для получения и хранения данных были использованы язык программирования PHP 7.2 и MySQL 5.6.

3.1. Данные по контрактам

По каждому контракту была получена следующая информация:

1) идентификационный номер налогоплательщика (ИНН) поставщика,
2) ИНН заказчика;
3) дата публикации на площадке государственных закупок,
4) регистрационный номер контракта (уникальный идентификатор закупки),
5) сумма контракта (итоговая сумма контракта после торгов),
6) дата начала действия контракта,

7) дата окончания действия контракта,
8) код по общероссийскому классификатору продукции по видам экономической деятельности (ОКПД) [1],
9) начальная максимальная цена контракта (начальная цена контракта на торгах),
10) причина расторжения контракта (если был расторгнут).

3.2. Данные по компаниям

В данных, полученных с торговой площадки «Сбербанк АСТ» по каждому контракту были получены ИНН заказчика и поставщика, по которым однозначно восстанавливаются компании. Для обучения нашей модели хорошо было бы получить как можно больше информации по этим компаниям. Есть несколько открытых источников, но в них не представлена какая-либо полезная информация для скоринга. Решено было использовать систему Seldon [12], так как с помощью неё можно получить большое количество данных по юридическим лицам.

Seldon — это сервис, предоставляющий информацию о юридических лицах. Он содержит много полезных данных о функционировании компаний: общую информацию, информацию о руководителе, финансовых показателях, госконтрактах и прочее. К сожалению, перечень данных по индивидуальным предпринимателям отличается от данных других юридических лиц, поэтому было решено исключить контракты, в которых одной из сторон был индивидуальный предприниматель. По ИНН, которые были получены вместе с информацией по контрактам, были программно скачаны данные по компаниям и записаны в локальную базу данных MySQL для дальнейшей обработки.

4. Признаки

Обычно для систем машинного обучения признаки берутся не только напрямую из данных, но и путём применения различных математических операций между несколькими параметрами. Благодаря этому можно найти новые признаки, которые способны улучшить точность модели. Следующие признаки были получены напрямую из собранных данных. Признаки, полученные математическими операциями из этих данных, будут представлены после.

4.1. Контракты

Из полученных данных были собраны следующие простые признаки, касающиеся непосредственно параметров контракта:

1) начальная максимальная цена контракта,
2) сумма контракта,

3) код ОКПД,
4) срок действия контракта в днях.

Начальная максимальная цена контракта — это сумма, с которой стартует аукцион. Далее заказчики предлагают меньшую цену с шагом не менее 0,5 % от цены.

Сумма контракта — это итоговая сумма, которая будет выплачена за исполнение контракта.

ОКПД — общероссийский классификатор продукции по видам экономической деятельности. Код ОКПД — это не более четырёх чисел, записанных через точку. Каждое число — номер подраздела спектра экономической деятельности. Например, код ОКПД 71.11.22.000 расшифровывается следующим образом:

- 71 — услуги в области архитектуры и инженерно-технического проектирования, технических испытаний, исследований и анализа,
- 71.11 — услуги в области архитектуры,
- 71.11.22 — услуги в области архитектуры, связанные с проектами строительства нежилых зданий и сооружений,
- 71.11.22.000 — услуги в области архитектуры, связанные с проектами строительства нежилых зданий и сооружений.

Срок действия контракта в днях — количество дней от даты начала действия контракта до даты окончания действия контракта.

4.2. Компании

Признаки, полученные из данных по контрактам и с помощью сервиса Seldon:

1) время существования компании до публикации контракта в днях,
2) количество компаний, в которых руководитель является текущим руководителем,
3) количество компаний, в которых руководитель является текущим совладельцем,
4) регион регистрации компании,
5) количество компаний по этому адресу,
6) адрес является адресом массовой регистрации (по данным ФНС),
7) индекс размера компании по данным Единого реестра малого и среднего предпринимательства,
8) наличие текущей задолженности перед ФНС по уплате налогов,
9) непредставления отчётности в ФНС более года,
10) логический признак включения юридического лица в реестр ФНС «Сведения о юридических лицах, связь с которыми по указанному ими адресу (месту нахождения), внесённому в Единый государственный реестр юридических лиц, отсутствует»,
11) численность сотрудников.

4.3. Подготовка признаков

Основная часть научного интереса работы легла на исследование признаков — выявление наиболее важных для предсказания дефолтности/успешности выполнения контракта. Для признаков, являющихся категориальными, было применено унитарное кодирование [11]. Далее был проведён поиск наиболее важных признаков путём генерации признаков, применением к ним различных математических операций.

Для генерации признаков были взяты основные математические операции (умножение, сложение, вычитание, деление) и применены к различным существующим признакам автоматическим перебором. В результате были взяты признаки, полученные с помощью применения функций одной переменной:

1) $p(x) = x^2$,
2) $p(x) = 1/x$,
3) $p(x) = \log(x)$,

где x — один из изначальных признаков, а p — новый признак, и признаки, полученные с помощью функций двух переменных:

1) $p(x, y) = x \times y$,
2) $p(x, y) = x/y$,
3) $p(x, y) = \frac{x-y}{y} \times 100\ \%$,

где x, y — изначальные признаки. При генерации не использовались категориальные признаки (к ним было применено унитарное кодирование). В результате из 23 признаков было дополнительно сгенерировано $23 \times 3 + 23 \times 23 \times 3 = 1656$ признаков. Перед тем как получить значимые признаки, найдём наиболее подходящие параметры классификатора.

5. Метод градиентного бустинга над решающими деревьями

Градиентный бустинг над решающими деревьями — один из самых эффективных методов машинного обучения [8]. У него есть ряд преимуществ:

— хорошо работает как с числовыми, так и с категориальными признаками,
— устойчив к переобучению,
— обучение эффективное и быстрое.

Бустинг над решающими деревьями — это метод построения решающих деревьев, при котором они строятся не независимо. Идея метода состоит в том, чтобы последовательно строить композицию решающих деревьев, каждое из которых ориентируется на примеры, которые предыдущие модели считали сложными и ошибочно классифицированными, стремится уменьшить значение функции потерь.

Как первое решающее дерево берётся дерево с заданной глубиной, определённым минимальным количеством листьев и коэффициентами, полученными после обучения на данных. Далее к текущей композиции (сумме) деревьев добавляется новое дерево так, чтобы у нового алгоритма было минимальное значение функции потерь. Это достигается путём перенастройки в листьях дерева с использованием градиентного спуска. Параметры алгоритма:

1) количество решающих деревьев, которые строятся в процессе обучения
2) глубина этих деревьев,
3) минимальное количество листьев,
4) скорость обучения — параметр, позволяющий управлять величиной коррекции весов на каждой итерации.

6. Поиск оптимальных параметров классификатора

В этой работе лучшим считается результат, который даёт наибольшую выгоду для банка. Мы посчитаем доход банка без использования этой модели и с её использованием. Затем посчитаем процент, на который данная система может увеличить этот доход.

6.1. Модель

Банк, выдавая банковские гарантии на контракт в случае успешного выполнения, получает в среднем 1 % от суммы контракта. В случае дефолта банк теряет 30 % от суммы контракта. Также необходимо учитывать тот факт, что в среднем дефолтный контракт в 2,6 раза дороже успешного (по статистике и данным, полученным от «Сбербанк АСТ»). В результате получаем, что потери от одного дефолтного контракта в 78 раз выше, чем потери дохода из-за отказа в выдаче банковской гарантии.

Таким образом, дефолт контракта в 78 раз хуже невыдачи успешного контракта. Поэтому для указания важности не пропустить дефолтный контракт из каждого такого контракта в обучающей выборке мы сделаем 78 копий. В результате получаем следующие выборки.

Успешных	Дефолтных	Дефолтных с учётом копий
Обучающая выборка		
624872	963	75114
Тестовая выборка		
154665	379	379

Далее подбираем параметры, которые покажут наибольшее увеличение дохода. Формула текущего дохода и его значение на тестовой выборке:

$$\text{Revenue}_0 = (\text{Contracts}_1 \cdot 0{,}01 - \text{Contracts}_0 \cdot 0{,}78) \cdot \text{PriceAv} \approx 1251 \cdot \text{PriceAv},$$

где Contracts_1 — количество успешных контрактов, Contracts_0 — количество дефолтных контрактов, PriceAv — средняя стоимость контракта. Подставляя в эту формулу вместо Contracts_1 количество верно отмеченных успешных контрактов и вместо Contracts_2 количество верно отмеченных дефолтных контрактов, получаем величину дохода Revenue_1, который получил бы банк при использовании разрабатываемой модели.

Далее изменение дохода будем рассчитывать по формуле

$$\Delta\text{Revenue} = \frac{\text{Revenue}_1 - \text{Revenue}_0}{\text{Revenue}_0} \cdot 100\ \%.$$

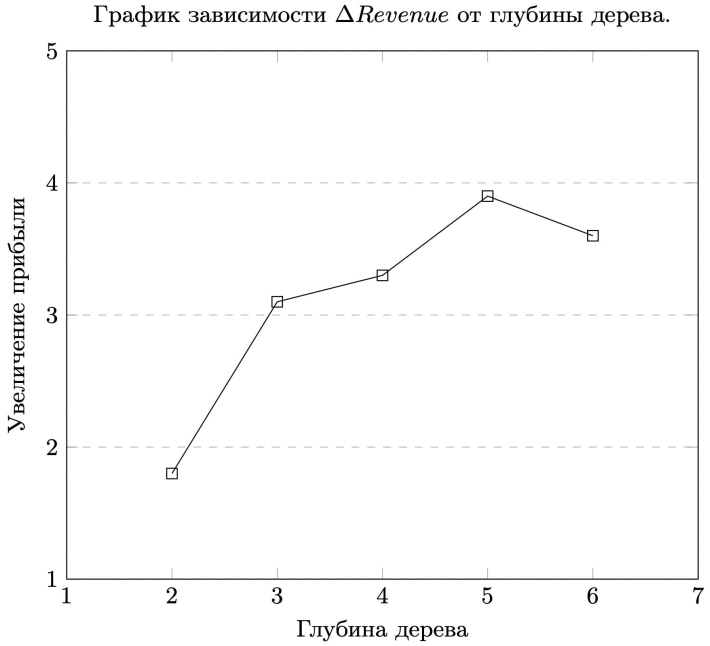
Относительно этого параметра будем искать наилучшие параметры классификатора. Для построения классификатора были использованы язык Python 3.6 и библиотека Scikit-learn. На изначальных данных, не используя сгенерированные признаки, найдём наилучшие параметры для классификатора.

Глубина дерева

Найдём наилучшее значения параметра «Глубина дерева». Подставим значения глубины дерева 2, 3, 4, 5, 6 при фиксированном количестве деревьев, равном 30, и скорости обучения 0,15. Значения 30 и 0,15 для начала выбираем стандартными для данного алгоритма.

Глубина дерева	2	3	4	5	6
Точность опред. усп. контр. на обучающей выборке	99,5 %	99,0 %	99,0 %	99,1 %	99,1 %
Точность опред. деф. контр. на обучающей выборке	14,6 %	24,8 %	36,6 %	48,9 %	54,3 %
Точность опред. усп. контр. на тестовой выборке	99,2 %	98,4 %	98,3 %	98,9 %	98,6 %
Точность опред. деф. контр. на тестовой выборке	11,9 %	21,6 %	22,7 %	22,2 %	22,7 %
ΔRevenue	1,8 %	3,1 %	3,3 %	3,9 %	3,6 %

Видим небольшой пик в значении 5. Вероятно, сначала при значениях 2, 3, 4 глубины недостаточно для хорошего обучения, а после 5 модель начинает переобучаться.

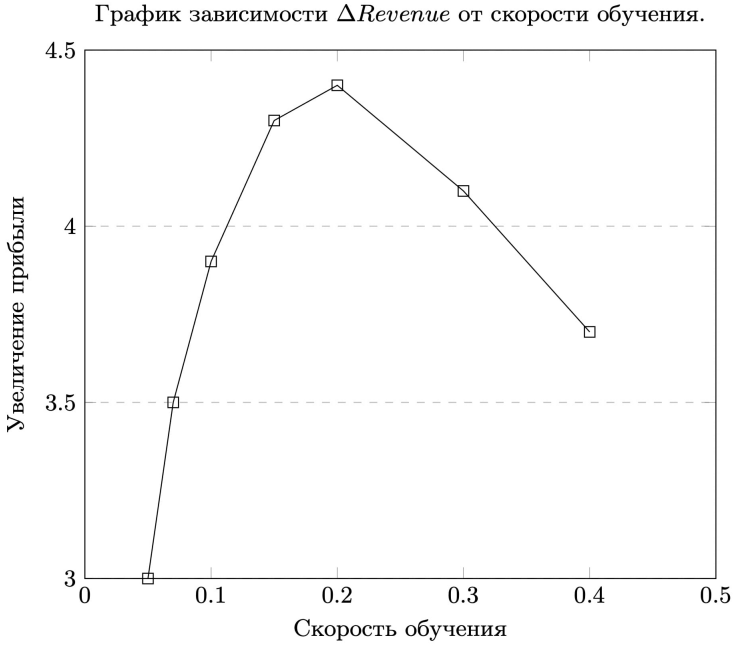


Скорость обучения

При фиксированных значениях количества деревьев 70 и глубины дерева 5 запускаем модель при следующих значениях скорости обучения: 0,05, 0,07, 0,09, 0,1, 0,12, 0,15, 0,2, 0,3, 0,4. Получаем следующие результаты.

Скорость обучения	0,05	0,07	0,1	0,15	0,2	0,3	0,4
Точность опред. усп. контр. на обучающей выборке	99,2 %	99,1 %	99,1 %	99,2 %	99,2 %	99,3 %	99,3 %
Точность опред. деф. контр. на обучающей выборке	40,1 %	49,1 %	57,4 %	66,5 %	72,4 %	83,3 %	90,0 %
Точность опред. усп. контр. на тестовой выборке	98,9 %	98,8 %	98,8 %	99,0 %	99,1 %	99,2 %	99,3 %
Точность опред. деф. контр. на тестовой выборке	18,7 %	21,1 %	23,0 %	23,0 %	23,7 %	21,4 %	19,3 %
ΔRevenue	3,0 %	3,5 %	3,9 %	4,3 %	4,4 %	4,1 %	3,7 %

Как видно из графика, наилучший результат находится рядом со значением 0,2. При слишком больших шагах получается маленькая точность, а при шагах, меньших 0,2, алгоритм начинает сходиться в каком-то локальном минимуме.



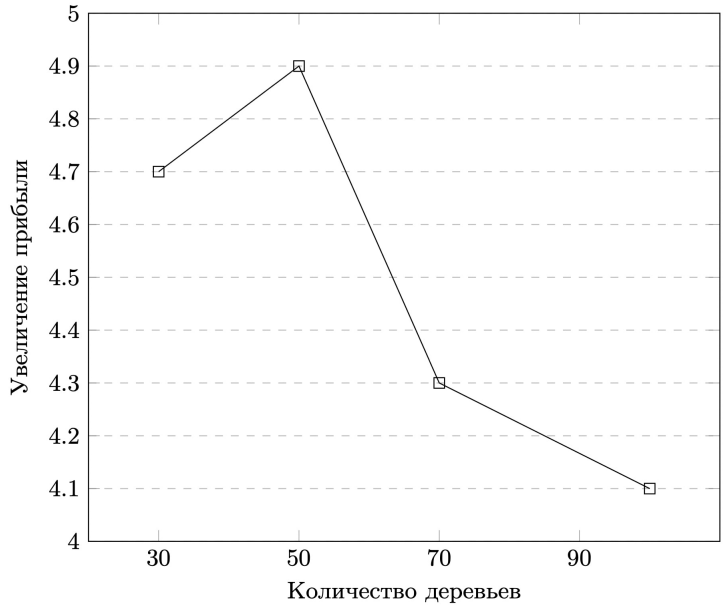
Количество деревьев

Теперь при фиксированных значениях глубины дерева и скорости обучения, равных 5 и 0,2 соответственно, рассмотрим, как меняется ΔRevenue при изменении количества деревьев.

Количество деревьев	30	50	70	100
Точность опред. усп. контр. на обучающей выборке	98,9 %	99,1 %	99,2 %	99,3 %
Точность опред. деф. контр. на обучающей выборке	56,1 %	64,5 %	72,4 %	78,9 %
Точность опред. усп. контр. на тестовой выборке	98,7 %	98,9 %	99,1 %	99,3 %
Точность опред. деф. контр. на тестовой выборке	26,5 %	26,7 %	23,0 %	21,1 %
ΔRevenue	4,7 %	4,9 %	4,3 %	4,1 %

Из таблицы и картинки видно понижение точности после 50 деревьев из-за переобучения.

График зависимости $\Delta Revenue$ от количества деревьев.



6.2. Итоговые параметры

Приведём итоговые параметры и результат работы модели на них.

Количество деревьев	Глубина дерева	Скорость обучения	% на усп.	% на деф.	ΔRevenue
50	5	0,2	98,9 %	26,7 %	4,9 %

7. Результат

7.1. Наиболее значимые признаки

Используя подобранные параметры классификатора и функцию SelectKBest из библиотеки scikit-learn, можно получить список наиболее значимых признаков. Этим путём были найдены следующие важные признаки, положительно влияющие на точность модели.

Процент падения цены контракта

Это абсолютная и относительная величина падения цены после завершения аукциона:

$$\text{PriceDecr} = \frac{\text{MaxPrice} - \text{Price}}{\text{MaxPrice}} \cdot 100\ \%,$$

где MaxPrice — начальная максимальная цена контракта, Price — цена контракта. Было выявлено, что вероятность дефолтности контракта возрастает с возрастанием процента падения цены. Это связано с тем, что изначально цена (начальная максимальная цена) находится на уровне рыночной, и если компания указала, что сможет выполнить контракт за цену сильно меньше рыночной, то повышается вероятность того, что она с этим не справится.

Стоимость контракта в день

$$\text{PricePerDay} = \frac{\text{Price}}{\text{Duration}},$$

где Duration — продолжительность действия контракта в днях. При слишком низкой или, наоборот, аномально высокой цене контракта в день вероятность дефолта растёт.

Цена контракта на одного работника

$$\text{PricePerEmployer} = \frac{\text{Price}}{\text{Employers}},$$

где Employers — количество работников у исполнителя по данным ФНС. Если это значение слишком велико, то компания может не справиться с контрактом. Если мало, то для компании может оказаться невыгодно выполнять контракт или же выполнение будет ненадлежащего качества и он окажется дефолтным.

Цена контракта, умноженная на изменение цены

$$\text{Price} \cdot \text{PriceDesr}.$$

Это ещё один признак, который значим для обучения. Вот только почему он оказался важен — не ясно.

Итак, из изначальных признаков мы сгенерировали новые, с помощью метода SelectKBest выбрали из них наиболее значимые для классификатора. Далее из сгенерированных оставляем только эти важные признаки и вместе с изначальными используем для дальнейшего обучения.

Значимость наиболее важных признаков	
Признак	Значимость
Процент падения цены контракта	0,24
Стоимость контракта в день	0,20
Продолжительность существования юр. лица исполнителя	0,12
Начальная максимальная цена	0,11
Продолжительность действия контракта	0,07
Цена контракта умноженная на изменение цены	0,07
Цена контракта на одного работника	0,06

7.2. Итоговая точность определения и прибыль

В итоге лучшее ΔRevenue при найденных выше параметрах и без сгенерированных признаков равно 4,9 %. Добавляя в обучающую выборку сгенерированные признаки, получаем ΔRevenue, равное 6,3 %.

Точность опред. усп. контр. на обучающей выборке	99 %
Точность опред. деф. контр. на обучающей выборке	68 %
Точность опред. усп. контр. на тестовой выборке	99 %
Точность опред. деф. контр. на тестовой выборке	31 %
ΔRevenue	6,3 %

8. Заключение

В исследовании была проведена работа по поиску наилучших параметров модели и наиболее значимых признаков для предсказания успешности выполнения контракта. Был рассмотрен один из самых эффективных методов классификации — градиентный бустинг над решающими деревьями.

Используя доступные данные, получилось увеличить прибыль на 6,3 %. Этот результат уже может быть полезен банкам. Если, к примеру, взять банк, который выдаёт банковские гарантии на контракты на общую сумму 3 млрд рублей в месяц и получает с этих гарантий 30 млн рублей дохода, то использование данной системы увеличит доход банка на 1,9 млн рублей в месяц.

Текущую модель можно и дальше улучшать, находя лучшие параметры, добавляя новые данные и генерируя новые признаки. К примеру, можно добавить информацию об завершённых контрактах исполнителя и его финансовые показатели, информацию о прошлых контрактах заказчика и так далее, чтобы достичь лучшей точности предсказания успешности выполнения контракта и, соответственно, увеличения прибыли банка.

Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект».

Литература

[1] Административно-управленческий портал. — http://www.aup.ru/okpd/.

[2] Гражданский кодекс РФ. Статья 368. Понятие и форма независимой гарантии.

[3] Дьяконов И. Д., Новикова С. В. Решение задачи прогнозирования при помощи градиентного бустинга над решающими деревьями // XV Международная научно-практическая конференция «Научный форум: технические и физико-математические науки». — https://nauchforum.ru/conf/tech/xv/35400.

[4] Открытый курс машинного обучения. Тема 10. Градиентный бустинг. — https://habr.com/ru/company/ods/blog/327250/.

[5] Сбербанк-АСТ — электронная торговая площадка. — https://www.sberbank-ast.ru.

[6] Федеральный закон от 05.04.2013 № 44-ФЗ (ред. от 01.05.2019) «О контрактной системе в сфере закупок товаров, работ, услуг для обеспечения государственных и муниципальных нужд».

[7] Ханжин С. В. Математическая модель кредитного скоринга потенциальных клиентов банка // Фундамент. и приклад. исслед.: проблемы и результаты: сб. матер. XX Междунар. науч.-практ. конф. Новосибирск, 2015. — Новосибирск: Изд-во ЦРНС, 2015. — С. 184—188. — https://cyberleninka.ru/article/n/matematicheskaya-model-kreditnogo-skoringa-potentsialnyh-klientov-banka.pdf.

[8] Freund Y., Schapire R. Experiments with a new boosting algorithm // Machine Learning. Proc. of the Thirteenth Int. Conf., 1996.

[9] Friedman J. Greedy function approximation: A gradient boosting machine. IMS 1999. Reitz Lecture. — https://jerryfriedman.su.domains/ftp/trebst.pdf.

[10] Mitchell T. Machine Learning. — McGraw-Hill, 1997.

[11] One-hot. — https://academy.yandex.ru/handbook/ml/article/linear-models.

[12] Seldon. Basis. — https://basis.myseldon.com/ru/landing.