УДК 338.2 **DOI: 10.14529/cmse200305**

ПРОГНОЗИРОВАНИЕ БАНКРОТСТВ ПРЕДПРИЯТИЙ С ПОМОЩЬЮ ЭКСТРЕМАЛЬНОГО ГРАДИЕНТНОГО БУСТИНГА

© 2020 В.В. Мокеев, Р.В. Войтецкий
Южно-Уральский государственный университет
(454080 Челябинск, пр. им. В.И. Ленина, д. 76)
E-mail: mokeyev@mail.ru, romanvoitetckii@gmail.com
Поступила в редакцию: 27.07.2020

Использование моделей прогнозирования банкротства предприятий для управления инвестиционными рисками лежит в основе управленческой деятельности финансовых учреждений. Важным фактором, позволяющим финансовым учреждениям определять объем капитала для покрытия кредитных потерь, является точность прогноза. В большинстве исследований для построения моделей банкротства предприятий используются традиционные методы статистики (например, дискриминантный анализ и логистическая регрессия). Однако точность построенных моделей обычно является достаточно низкой. Это обусловлено несбалансированностью классов обучающих выборок (доля фирм-банкротов составляет несколько процентов от общего числа фирм), которые используются при построении моделей. В настоящее время широкое распространение получают такие методы машинного обучения как метод случайного леса и метод градиентного бустинга. В данном исследовании основной акцент делается на использовании экстремального градиентного бустинга для прогнозирования банкротства. Экстремальный градиентный бустинг, используя LASSO или Ridge регуляризацию, штрафует сложные модели, что помогает избежать переобучения. Также в ходе обучения экстремальный градиентный бустинг заполняет пропущенные значения в наборе данных в зависимости от величины потерь. В статье для повышения эффективности экстремального градиентного бустинга предлагается использовать технологию SMOTE для улучшения сбалансированности классов. Метрики качества решений, полученных улучшенным экстремальным градиентным бустингом, сравниваются с решениями полученными другими методами.

Ключевые слова: экстремальный градиентный бустинг, банкротство, предприятие, synthetic minority oversampling technique.

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Мокеев В.В., Войтецкий Р.В. Прогнозирование банкротств предприятий с помощью экстремального градиентного бустинга // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2020. Т. 9, № 3. С. 77–90. DOI: 10.14529/cmse200305.

Введение

Инвестиционные риски являются основной проблемой для финансовых учреждений, что заставляет их проверять и контролировать финансовую платежеспособность предприятия. Модели банкротств используются финансовыми учреждениями для оценки кредитного убытка, который они могут понести, если их контрагенты не возместят свои долги. Точность прогноза является ключевым фактором, поскольку она позволяет финансовым учреждениям определять объем капитала, необходимый для покрытия кредитных потерь. В настоящее время существует достаточно большое число исследований, направленных на повышение точности моделей банкротств предприятий. Практическая значимость прогноза банкротства связана со значительным интересом со стороны кредиторов в точном прогнозировании будущего существования компаний.

В большинстве исследований прогноз банкротства рассматривается как проблема бинарной классификации. Целевая переменная моделей обычно принимает два значения: 0 (платежеспособное предприятие) и 1 (предприятие банкрот). Традиционно для оценки банкротства используются дискриминантный анализ [1–5] и регрессионный ана-

лиз [5–7]. Альтман использовал линейный дискриминантный анализ (Linear Discriminant Analysis, LDA) для того, чтобы различать предприятия банкротов от не банкротов [1]. LDA применяет линейную комбинацию показателей для определения рейтинга предприятия. Эта оценка затем применяется для разделения предприятий на банкротов и не банкротов. В работе [2] LDA используется для прогнозирования риска банкротства российских предприятий. Применимость LDA и квадратичного дискриминантного анализа исследуется в работах [4, 5]. В работе [5] модели банкротства предприятий строятся с использованием логистической регрессии, но, в отличие от модели Альтмана [1], модель Олсона [5] определяет вероятность банкротства. Несмотря на относительную легкость построения моделей с помощью дискриминантного анализа и логистической регрессии, использование этих моделей показывает плохую способность к обобщению и низкую точность прогнозирования [8]. Поэтому в дальнейшем многие исследователи используют методы машинного обучения, такие как нейронные сети [9–12], метод опорных векторов [13, 14], бустинговые методы [15, 16].

Одной из характерных особенностей задачи прогнозирования банкротств является то, что доля фирм-банкротов составляет несколько процентов от общего числа фирм. Поэтому исследователи сталкиваются с несбалансированными наборами данных, в которых число обанкротившихся компаний явно превосходит число необанкротившихся компаний. В ряде работ предпринимаются попытки найти методы улучшения несбалансированных наборов данных. Большинство из них используют механизмы сбалансирования выборочных распределений. В работе [17] используется несколько методов, чтобы улучшить два сильно несбалансированных набора данных. Эти методы позволяют методам прогнозирования банкротства достигать лучших результатов, чем прогнозы по исходным несбалансированным наборам данных. В работе [18] используют обучающую выборку, которая содержит 620 обанкротившихся образцов и 7398 необанкротившихся фирм. Для повышения точности прогнозирования банкротств предлагается гибридный метод недостаточной выборки, который сочетает в себе метод ближайших соседей и метод опорных векторов. Результаты исследований показывают, что предлагаемый гибридный метод повышает точность классификаторов, таких как логистическая регрессия, дискриминантный анализ, дерево решений и машины опорных векторов. В работе [19] оцениваются четыре метода улучшения сбалансированности наборов данных: случайная избыточная выборка, случайная недостаточная выборка, метод искусственного увеличения меньшинства (SMOTE) и EasyEnsemble. Результаты исследований показывают, что SMOTE превосходит другие методы улучшения сбалансированности наборов данных в плане повышения точности прогнозирования.

Цель данного исследования заключается в исследовании эффективности метода экстремального бустинга в сочетании с методом улучшения сбалансированности обучающей выборки для решения задач прогнозирования банкротств предприятий. Предлагается сравнить полученные результаты с решениями, полученными другими методами.

Статья организована следующим образом. В разделе 1 описывается экстремальный градиентный бустинг, а также технология SMOTE для улучшения сбалансированности наборов данных. В разделе 2 описываются наборы данных, содержащие финансовые показатели предприятий и метку, указывающую на статус банкротства предприятий, а также результаты исследования эффективности моделей прогнозирования банкротств. В заключении приводится краткая сводка результатов, полученных в работе, и указаны направления дальнейших исследований.

1. Теоретическая основа

1.1. Экстремальный градиентный бустинг

Экстремальный градиентный бустинг (Extreme Gradient Boosting, XGB) представляет развитие метода градиентного бустинга [20, 21]. При обучении с учителем набор данных $D = \{(x_i, y_i): x_i \in R^n, y_i \in R\}$, состоит из n объектов с m признаками и n метками. Необходимо построить модель, которая как можно более точно сможет предсказывать метки для каждого нового объекта. Бустинговая модель F использует N аддитивных функций $f_i(x)$ для прогнозирования меток

$$\tilde{y}_i = F(x_i) = \sum_{j=1}^{N} f_j(x_i). \tag{1}$$

Для обучения набора функций мы минимизируем функцию потерь

$$L(F) = \sum_{i=1}^{n} l(y_i, \tilde{y}_i) + \sum_{j=1}^{K} R(f_j), \tag{2}$$

где $R(f_j)$ является коэффициентом регуляризации, который ограничивает сложность модели. Функция потерь $L(F)$ содержит K функций в качестве параметров, поэтому ее очень трудно оптимизировать напрямую. Вместо этого мы оптимизируем модель аддитивно. Пусть $\tilde{y}_i^t$ будет предсказанием i-го образца на t-й итерации. Мы добавим f_t, чтобы минимизировать

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \tilde{y}_i^{(t-1)} + f_t(x_i)) + R(f_t). \tag{3}$$

Это означает, что f_t добавляется так, чтобы максимально улучшить нашу модель для каждой итерации. Мы используем приближение второго порядка, которое использует градиент этой промежуточной функции потерь $L^{(t)}$. По этой причине мы называем его алгоритмом градиентного бустинга.

Основные отличия экстремального градиентного бустинга от градиентного бустинга связаны с регуляризацией и работой с пропущенными данными. Экстремальный градиентный бустинг штрафует сложные модели, применяя Ridge-регуляризацию или регуляризацию LASSO, что позволяет снизить риски переобучения. Также в ходе обучения экстремальный градиентный бустинг использует алгоритм заполнения пропущенных значений в зависимости от величины потерь.

Точность решений получаемых с помощью XGB зависит от гиперпараметров метода. Важнейшими из них являются максимальная глубина дерева, скорость обучения, число деревьев. Увеличение максимальной глубины дерева делает модель более сложной и повышает вероятность переобучения. Более высокая скорость обучения приводит к тому, что каждое дерево вносит более серьезные поправки в решение. Это приводит к усложнению модели и повышает вероятность переобучения. Увеличение числа деревьев также повышает сложность модели, но при этом появляется возможность повысить точность получаемых решений.

1.2. Экстремальный градиентный бустинг и метод искусственного увеличения экземпляров миноритарного класса

Для повышения качества прогнозирования банкротств предлагается использовать при построении моделей методом экстремального градиентного бустинга метод улучшения сбалансированности обучающей выборки. В качестве последнего предлагается использовать метод Synthetic Minority Oversampling Technique (SMOTE), который основан на идее генерации некоторого количества искусственных образцов, «похожих» на имеющиеся в классе банкротов, но не дублирующих их. Для создания нового образца находят разность $d = X_b - X_a$, где X_a, X_b — это векторы признаков «соседних» образцов a и b класса предприятий банкротов. Их находят, используя алгоритм ближайшего соседа. В нашем случае необходимо и достаточно для генерирования нового образца получить набор из k ближайших соседей, из которого будут выбраны объекты a и b.

Далее вектор разности нового образца $\tilde{d}$ получается путем умножения каждого элемента вектора d на случайное число в интервале (0, 1). Вектор признаков нового образца вычисляется путем сложения $X_c = X_a + \tilde{d}$. Метод SMOTE позволяет задавать количество новых образцов, которое необходимо искусственно сгенерировать.

Для обеспечения контроля качества обучения моделей предлагается следующая схема построения моделей. Весь набор данных предварительно делится на обучающий и тестовый набор. Для обучения модели используется процедура кросс-валидации, в рамках которой набор делится на K блоков (folds). Процесс обучения выполняется K раз на разных обучающих выборках, состоящих из $K–1$ блоков. Для улучшения сбалансированности классов обучающих выборок они расширяются путем генерации искусственных образцов, «похожих» на имеющиеся в классе банкротов, но не дублирующих их. Для генерации используется метод SMOTE. Для контроля качества обучения используется оставшийся валидационный блок a. Таким образом, мы получаем K моделей и K валидационных блоков, которые образуют валидационный набор. Валидационный набор используется для оценки качества обучения. Для оценки обобщающей способности модели используется тестовый набор. Поскольку в рамках процедуры кросс-валидации было построено K моделей, мы получаем на тестовом наборе K прогнозов, которые затем усредняются.

Таким образом, метод SMOTE органично встраивается в схему обучения моделей методом экстремального градиентного бустинга и может представлять улучшенную версию экстремального градиентного бустинга.

1.3. Метрика качества

Традиционно для оценки качества моделей классификации используется метрика *accuracy*, которая численно равна доли правильно классифицированных объектов. Однако для несбалансированных классов такая метрика является не очень удобной. Для оценки качества модели на каждом из классов по отдельности используются метрики *precision* (точность) и *recall* (полнота). *Precision* можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а *recall* показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм. *Precision* и *recall* не зависят, в отличие от *accuracy*, от соотношения классов и потому применимы в условиях несбалансированных наборов.

Обычно при выборе гипер-параметров метода используется одна метрика, улучшение которой мы и ожидаем увидеть на тестовом наборе. В качестве такой метрики

предлагается использовать метрику *F1 score*, которая представляет среднее гармоническое *precision* и *recall*. Для вычисления метрик precision, *recall*, *F1* и *accuracy* требуется преобразование вероятностного результата в бинарную метку, т.е. необходимо выбрать какой-либо порог, при котором результат становится 0 или 1. Естественным и близким является порог, равный 0,5, но он не всегда оказывается оптимальным, особенно при отсутствии баланса классов. Одним из способов оценки модели в целом, не привязываясь к конкретному порогу, является метрика AUC, которая численно равна площади под кривой ошибок.

2. Экспериментальное исследование

2.1. Набор данных

В данной статье данные были взяты с сайта UCI machine learning repository[1]. Изначально они были извлечены с ресурса Emerging Markets Information Services[2], который представляет собой базу данных, содержащую информацию о развивающихся рынках по всему миру. Обанкротившиеся компании были проанализированы в период 2000–2012 годов, в то время как все еще действующие компании были оценены с 2007 по 2013 год. Всего предоставлено пять наборов данных, которые содержат финансовые показатели предприятий и метку, которая указывает на статус банкротства через 1, 2, 3, 4 года и 5 лет. Метаданные наборов данных представлены в табл. 1.

Таблица 1

Основные характеристики набора данных

Характеристика	Метка статуса банкротств				
	1 год	2 года	3 года	4 года	5 лет
Число предприятий	5910	9792	10503	10173	7027
Доля предприятий банкротов (%)	6,93	5,26	4,71	3,93	3,85
Доля пропущенных значений (%)	1,21	1,37	1,44	1,83	1,27

Таким образом, мы имеем 5 наборов, которые не являются равноценными. Три набора с метками статуса банкротств через 2, 3 и 4 года содержат близкое число предприятий примерно около 10000. Набор данных с метками статуса банкротств через 1 год имеет на 40% меньше предприятий, а набор данных с метками статуса банкротства через 5 лет включает на 30% меньше предприятий, чем наборы данных с метками статуса банкротств через 2, 3 и 4 года. В наборах каждое предприятие описывается 64 финансовыми показателями, которые измеряют рентабельность, активы предприятий и их денежные потоки. Полное описание показателей можно найти в работе [15], в которой исследуются эффективность методов машинного обучения при построении моделей прогнозирования банкротств. Однако исследования выполненные в работе обладают рядом недостатков.

Во-первых, для процедуры кросс-валидации используется весь набор, и качество моделей определяется на валидационном наборе. Такая оценка характеризует только качество обучения модели, особенно при использовании таких методов как метод случайного леса (Random Forest) и экстремальный градиентный бустинг (Extreme Gradient Boosting) [15]. Поскольку важно знать, насколько хорошо модель обобщает результат на новых, ранее неизвестных данных, кроме валидационного набора требуется отложенный

[1] https://www.re3data.org/repository/r3d100010960

[2] https://www.emis.com/

тестовый набор, который в обучении не принимает участие. Во-вторых, в качестве метрики качества используется метрика AUC, которой не достаточно для полной оценки качества прогнозирования банкротства предприятий. Поэтому мы будем вместе с AUC использовать метрику $F1$ для каждого класса.

2.2. Предварительная обработка данных

При построении моделей используются наборы данных, описанные выше, которые делятся следующим образом. Весь набор делится на две части: обучающий (70%) и тестовый (30%) наборы. Выделение достаточно большого тестового набора позволяет повысить достоверность оценки обобщающей способности модели.

Одна из первых задач предварительной обработки данных заключается в заполнении пропущенные значения. Существует несколько стратегий заполнения пропущенных данных. Наиболее популярными стратегиями являются стратегии, базирующиеся на присвоении пропущенным значениям средних (mean) или медианных (median) значений признаков. При использовании деревьев решений эффективной является стратегия маркировки пропущенных значений, например, присвоение пропущенным значениям больших отрицательных чисел (constant).

Таблица 2 представляет результаты сравнения различных стратегий заполнения пропущенных значений для набора данных с меткой статуса банкротства 5 лет. Метрика качества моделей, полученных методом XGB. Столбцы таблиц с заголовками «N» содержат значения метрик $F1$ и AUC для предприятий, не являющихся банкротами, а столбцы, озаглавленные «B», — метрик $F1$ и AUC для предприятий-банкротов.

Таблица 2

Сравнение качества решений для различных стратегий заполнения пропущенных значений

Стратегия	$F1$		AUC	$F1$		AUC
	N	B		N	B	
Mean	0,9901	0,6595	0,943	0,9887	0,6515	0,962
Median	0,9890	0,6067	0,937	0,9883	0,6308	0,964
Constant	0,9907	0,6855	0,941	0,9904	0,7153	0,969

Анализ результатов, представленных в таблицах, показывает стратегия constant, которая заполняет пропущенные значения постоянными числами, равными -99999, является наиболее эффективной. Исследования на других наборах данных также показывает небольшое преимущество стратегии constant. Исследования, проведенные на других наборах данных, подтверждают предпочтение стратегии constant перед другими стратегиями.

Вторая задача, которая решается на этапе подготовки данных, заключается в выборе наиболее информативных признаков из общего числа. Наличие в данных неинформативных признаков приводит к переобучению модели, снижению точности получаемых решений и росту времени обучения модели. В табл. 3 представлено сравнение качества моделей, полученных при различных количествах признаков, которые были отсортированы по коэффициенту значимости.

Таблица 3

Сравнение качества решений для различных комбинаций признаков

Число признаков	*F1*		*AUC*	*F1*		*AUC*
	N	*B*		*N*	*B*	
64	0,9907	0,6855	0,941	0,990	0,7153	0,969
50	0,9909	0,6926	0,9491	0,99	0,6953	0,970
40	0,9909	0,6947	0,9521	0,991	0,7338	0,970
30	0,9909	0,6947	0,9459	0,991	0,7230	0,941

В топ-10 наиболее важных признаков входят следующие признаки:

1) отношение прибыли от операционной деятельности к финансовым расходам;
2) коэффициент текущей ликвидности;
3) отношение операционных расходов к суммарным обязательствам;
4) логарифм от суммарных активов;
5) отношение суммы денежных средств, краткосрочных инвестиций, дебиторской задолженности минус краткосрочные обязательства к разности операционных расходов и амортизации;
6) отношение суммарных издержек к суммарным продажам;
7) отношение выручки от продаж к суммарным активам;
8) отношение дебиторской задолженность умноженной на 360 к выручке от продаж;
9) отношение разности текущих активов и запасов к долгосрочным обязательствам;
10) отношение суммы валовой прибыли, иных активов и транзакции, финансовых затрат к суммарным активам.

Эти показатели измеряют не только рентабельность предприятий, но и активы предприятий, и их денежные потоки. Анализируя результаты выполненных исследований можно сказать, что эти показателя появляются в списке важных показателей для каждого набора данных. Помимо этих 10 показателей были отобраны еще 35 показателей, которые положительно влияют на качество моделей.

2.3. Сравнение качества моделей, построенных различными методами

Для построения моделей используются следующие методы:

– Linear Discriminant Analysis (LDA);
– Logistic Regression (LR);
– Random Forest (RF);
– Extreme gradient boosting (XGB);
– Extreme gradient boosting and Smote (XGB-Sm).

При использовании методов LR, RF, XGB, XGB-Sm осуществлялся поиск наилучших гипер-параметров путем перебора их значений в заданных диапазонах. В ходе исследования было проведено 10 экспериментов. В ходе эксперимента каждый набор делится на обучающий (70%) и тестовый (30%) наборы. При построении модели на обучающем наборе используется процедура кросс-валидации на 10-ти блоках. В результате качество модели оценивается на валидационном и тестовом наборах. В процессе оценки качества модели с использованием метрик *F1* и *accuracy*, выбор порога, при котором вероятностный результат преобразовывается в бинарную метку, осуществляется из условия максимизации метрики *F1* для класса предприятий банкротов. Метрики качества, полученные в ходе 10 экспериментов, усредняются. В табл. 4–8 представлена метрика качества моделей, полученных различными методами: среднее значение метрики

F1 для действующих предприятий (N) и предприятий банкротов (B), и среднее значение метрики AUC и *accuracy*.

Таблица 4

Сравнение качества моделей с горизонтом прогнозирования 1 год

Метод	Валидационный набор				Тестовый набор			
	F1		AUC	Acc	*F1*		AUC	Acc
	N	B			N	B		
LDA	0,9662	0,5075	0,855	0,9367	0,9666	0,5075	0,855	0,9374
LR	0,9713	0,4851	0,709	0,9456	0,9709	0,319	0,716	0,9447
RF	0,9638	0,5767	0,919	0,9333	0,9669	0,6197	0,935	0,9391
XGB	0,9813	0,6971	0,953	0,9647	0,9830	0,7136	0,968	0,9679
XGB-Sm	0,9814	0,7252	0,957	0,9652	0,9829	0,7534	0,968	0,9690

В табл. 4 представлены результаты, полученные на наборе данных с метками статуса банкротства через 1 год. Доля предприятий банкротов в валидационном и тестовом наборах составляют, соответственно, 0,0696 и 0,0688. Таким образом, если в качестве решения использовать простое решение (все предприятия являются действующими), то значения метрики *accuracy* будут равны 0,9304 и 0,9312, соответственно. Будем считать эти значения пороговыми, т.е. прогнозы, метрика *accuracy* которых лежит ниже пороговых значений, будем считать плохими. Как видно из таблицы, у всех моделей значение метрики качества *accuracy* превышает их пороговые значения. Качество моделей, полученных методами LDA и RF, превышает пороговые значения незначительно, в то время как метрика AUC имеет достаточно высокие значения. Использование метода логистической регрессии позволяет получить прогнозы, метрика *accuracy* которых превышает пороговые значения почти на 1,5%, а метрика AUC намного ниже аналогичной метрики в случае применения методов LDA и RF.

При использовании логистической регрессии построенные модели позволяют точно классифицировать 106 из 288 и 39 из 122 предприятий банкротов для валидационного и тестового наборов, соответственно. Также правильно классифицируются 3806 из 3849 и 1636 из 1651 действующих предприятий для валидационного и тестового наборов данных, соответственно. При использовании моделей, построенных с помощью LDA, правильно классифицируются 135 из 288 и 57 из 122 предприятий банкротов для валидационного и тестового наборов, соответственно. Из действующих предприятий валидационного и тестового наборов правильно классифицируются 3750 из 3849 и 1605 из 1651, соответственно.

Модели, построенные с помощью RF, позволяют правильно классифицировать 188 из 288 и 88 из 122 предприятий банкротов валидационного и тестового наборов, соответственно. Из действующих предприятий валидационного и тестового наборов правильно классифицируются 3673 из 3849 и 1577 из 1651, соответственно.

Наибольшую точность демонстрируют модели, полученные с помощью XGB и XGB-Sm. Улучшение сбалансированности обучающих наборов данных повышают качество прогноза, так метрика качества *F1* для предприятий банкротов увеличивается на 2,8% и 3,9% для валидационного и тестового наборов данных, соответственно. Метрика *F1* для действующих предприятий практически не изменяется. Оптимальный коэффициент баланса составляет 0,09.

Таблица 5

Сравнение качества моделей с горизонтом прогнозирования 2 года

Метод	Валидационный набор				Тестовый набор			
	F1		*AUC*	*Accuracy*	*F1*		*AUC*	*Accuracy*
	N	*B*			*N*	*B*		
LDA	0,9753	0,5038	0,825	0,9529	0,9732	0,4526	0,825	0,9489
LR	0,9742	0,4329	0,690	0,9507	0,9743	0,3777	0,658	0,9506
RF	0,9821	0,6274	0,883	0,9610	0,9819	0,5887	0,886	0,9632
XGB	0,9839	0,6201	0,934	0,9691	0,9835	0,5903	0,934	0,9683
XGB-Sm	0,9835	0,6861	0,942	0,9710	0,9823	0,6374	0,944	0,9694

В табл. 5 представлена метрика качества моделей, полученных на наборах с метками статуса банкротства предприятий через 2 года. Пороговые значения метрики *accuracy* для валидационного и тестового наборов составляют 94,61% и 95,03%, соответственно. Как видно из таблицы, на тестовом наборе модели LDA имеют значение *accuracy* ниже порогового значения, а значение метрики *accuracy* моделей LR незначительного превышает соответствующее пороговое значение.

Модели, построенные методами XGB и XGB-Sm, показывают наиболее высокие значения метрик качества. Улучшение сбалансированности обучающих наборов данных повышает качество прогноза, например, метрика качества *F1* для предприятий банкротов увеличивается на 6,6% и 4,7% для валидационного и тестового наборов, соответственно. Хотя при этом метрика *F1* для действующих предприятий немного снижается (0,04% и 0,01%). Оптимальный коэффициент баланса классов составляет 0,09.

Таблица 6

Сравнение качества моделей с горизонтом прогнозирования 3 года

Метод	Валидационный набор				Тестовый набор			
	F1		*AUC*	*Accuracy*	*F1*		*AUC*	*Accuracy*
	N	*B*			*N*	*B*		
LDA	0,9526	0,3421	0,790	0,9116	0,9543	0,3182	0,790	0,9143
LR	0,9677	0,2720	0,661	0,9381	0,9650	0,2327	0,663	0,9330
RF	0,9795	0,5207	0,885	0,9607	0,9752	0,4186	0,856	0,9524
XGB	0,9842	0,5591	0,931	0,9695	0,9828	0,4274	0,912	0,9667
XGB-Sm	0,9852	0,6534	0,938	0,9716	0,9823	0,5597	0,919	0,9660

В табл. 6 представлена метрика качества моделей, полученных на наборах с метками статуса банкротства предприятий через 3 года. Пороговые значения метрики *accuracy* для валидационного и тестового наборов составляют 95,25% и 95,36%, соответственно. Как видно из таблицы, метрика *accuracy* моделей, построенных методами LDA и LR, на тестовом наборе лежит ниже пороговых значений. Модели, построенные с использованием метода RF, при классификации предприятий тестового набора показывают результаты, в которых доля правильно классифицированных предприятий ниже порогового значения, что также говорит о невысоком качестве модели.

Модели, построенные методами XGB и XGB-Sm, показывают наиболее высокие значения метрики качества. Улучшение сбалансированности набора данных повышают качество прогноза, Метрика качества *F1* для предприятий банкротов увеличивается на 9,5% и 13,1% на валидационном и тестовом наборах данных, соответственно. Хотя при

этом метрика *F1* для действующих предприятий увеличивается на 0,1% на валидационной наборе и снижается на 0,05% на тестовом наборе. Оптимальный коэффициент баланса классов составляет 0,06.

Таблица 7

Сравнение качества моделей с горизонтом прогнозирования 4 года

Метод	Валидационный набор				Тестовый набор			
	F1		*AUC*	*Accuracy*	*F1*		*AUC*	*Accuracy*
	N	*B*			*N*	*B*		
LDA	0,9663	0,3289	0,721	0,9358	0,9693	0,2800	0,721	0,9410
LR	0,9741	0,2452	0,601	0,9499	0,9809	0,2083	0,606	0,9626
RF	0,9822	0,5243	0,865	0,9656	0,9855	0,5275	0,850	0,9718
XGB	0,9868	0,5667	0,914	0,9744	0,9882	0,5333	0,937	0,9771
XGB-Sm	0,9874	0,6547	0,921	0,9757	0,9882	0,6316	0,939	0,9780

В табл. 7 представлена метрика качества моделей, полученных на наборах с метками статуса банкротства предприятий через 4 года. Пороговые значения метрики *accuracy* для валидационного и тестового наборов данных составляют 95,87% и 96,53%, соответственно. Как видно из таблицы, метрика качества *accuracy* моделей, построенных методами LDA и LR, на валидационном и тестовом наборах имеют значение ниже пороговых значений.

Модели, построенные методами XGB и XGB-Sm, показывают наиболее высокие значения метрики качества. Улучшение сбалансированности набора данных повышают качество прогноза, так метрика качества *F1* для предприятий банкротов увеличивается на 8,8% и 9,8% для валидационного и тестового наборов, соответственно. Хотя при этом метрика *F1* для действующих предприятий увеличивается на 0,06% на валидационном наборе и не меняется на тестовом наборе. Оптимальный коэффициент баланса классов составляет 0,06.

Таблица 8

Сравнение качества моделей с горизонтом прогнозирования 5 лет

Метод	Валидационный набор				Тестовый набор			
	F1		*AUC*	*Accuracy*	*F1*		*AUC*	*Accuracy*
	N	*B*			*N*	*B*		
LDA	0,9775	0,4928	0,842	0,9569	0,9812	0,5870	0,924	0,9640
LR	0,9814	0,5294	0,882	0,9642	0,9820	0,5629	0,893	0,9654
RF	0,9815	0,4699	0,908	0,9642	0,9808	0,5185	0,942	0,9630
XGB	0,9909	0,6926	0,937	0,9823	0,9909	0,6947	0,942	0,9806
XGB-Sm	0,9910	0,7075	0,943	0,9827	0,9914	0,7552	0,969	0,9844

В табл. 8 представлена метрика качества моделей, полученных на наборах с метками статуса банкротства предприятий через 5 лет. Пороговые значения метрики *accuracy* для валидационного и тестового наборов, составляют 96,28% и 95,83%, соответственно. Как видно из таблицы, метод LR на валидационном наборе имеют значение *accuracy* ниже пороговых значений, хотя на тестовом наборе ситуация уже меняется.

Модели, построенные методами XGB и XGB-Sm, показывают наиболее высокие значения метрики качества. Улучшение сбалансированности обучающих наборов данных

повышают качество прогноза, но незначительно. Оптимальный коэффициент баланса классов составляет 0,07.

Таким образом, модели, построенные методами LDA и LR, дают самую низкую точность прогноза, которая часто оказвается ниже пороговых значений. Модели, построенные методом RF, хотя и дают более высокую прогноза, но в некоторых случаях не позволяет получить модели с хорошим качеством. Модели, построенные XGB-Sm, демонстриуют наиболее высокие значения метрики качеством как на валидационном, так и на тестовом наборах. Следует отметить, что в случае если число искусственных образцов, генерируемых методом SMOTE, превышает исходное число образцов, качество моделей, построенных экстремальных градиентным бустингом, существенно падает.

Заключение

В статье рассмотрена проблема прогнозирования банкротства на основе финансовых факторов. Для решения поставленной задачи классификации используется модель, построенная с помощью экстремального градиентного бустинга. Результаты, полученные экстремальным градиентным бустингом, были значительно лучше, чем результаты, полученные такими известными методами как линейный дискриминантный анализ, логистическая регрессия, которые широко применяются для прогнозирования финансового состояния компаний. В работе предложено расширение экстремального градиентного бустинга, которое улучшает дисбаланс классов обучающих наборов с помощью метода SMOTE. Применение такого подхода привело к улучшению качества прогнозирования.

Будущие исследования планируется вести в направлении учета темпов роста финансовых показателей при построении моделей, чтобы оценить влияние времени на вероятность банкротства предприятий.

Литература

1. Altman E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy // The Journal of Finance. 1968. Vol. 23, no. 4. P. 589–609. DOI: 10.1111/j.1540-6261.1968.tb00843.x.
2. Lugovskaya L. Predicting Default of Russian Smes on the Basis of Financial and Nonfinancial Variables // Journal of Financial Services Marketing. 2010. Vol. 14, no. 4. P. 301–313. DOI: 10.1057/fsm.2009.28.
3. Deakin E. A Discriminant Analysis of Predictors of Business Failure // Journal of Accounting Research. 1972. Vol. 10, no. 1. P. 167–179. DOI: 10.2307/2490225.
4. Antunesa F., Ribeiroa B., Pereirab F. Probabilistic Modeling and Visualization for Bankruptcy Prediction // Applied Soft Computing. 2017. Vol. 60. P. 831–843. DOI: 10.1016/j.asoc.2017.06.043.
5. Ohlson J.A. Financial Ratios and the Probabilistic Prediction of Bankruptcy // Journal Of Accounting Research. 1980. Vol. 18, no. 1. P. 109–131. DOI: 10.2307/2490395.
6. Martin D. Early Warning of Bank Failure: A Logit Regression Approach // Journal of Banking & Finance. 1977. Vol. 1, no. 3. P. 249–276.
7. Wiginton J.C. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior // Journal of Financial and Quantitative Analysis. 1980. Vol. 15. P. 757–770. DOI: 10.2307/2330408.
8. Begley J., Ming J., Watts S. Bankruptcy Classification Errors in the 1980s: an Empirical Analysis of Altman's and Ohlson's Models // Review of Accounting Studies. 1996. Vol. 1, no. 4. P. 267–284. DOI: 10.1007/bf00570833.

9. Wilson R.L, Sharda R., Bankruptcy Prediction Using Neural Networks // Decision Support Systems. 1994. Vol. 11, no. 5. P. 545–557. DOI: 10.1016/0167-9236(94)90024-8.
10. Tam K.Y., Kiang M.Y. Managerial Applications of Neural Networks: the Case of Bank Failure Predictions // Management Science. 1992. Vol. 38, no. 7. P. 926–947. DOI: 10.1287/mnsc.38.7.926.
11. Altman E.I., Marco G., Varetto F. Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience) // Journal of Banking & Finance. 1994. Vol. 18, no. 3. P. 505–529. DOI: 10.1016/0378-4266(94)90007-8.
12. Ciampi F., Vallini C., Gordini N. Using Artificial Neural Networks Analysis for Small Enterprise Default Prediction Modeling: Statistical Evidence from Italian Firms // Oxford Business & Economics Conference Proceedings, Association for Business and Economics Research, ABER. 2009. Vol. 1. P. 126.
13. Wei L., Li J., Chen Z. Credit Risk Evaluation Using Support Vector Machine with Mixture of Kernel // Proceedings of the 7th International Conference on Computational Science. Lecture Notes in Computational Science and Engineering. 2007. Vol. 4488. P. 431–438. DOI: 10.1007/978-3-540-72586-2_62.
14. Härdle W.K., Lee Y.J., Schäfer D. The Default Risk of Firms Examined with Smooth Support Vector Machines // Discussion Papers, German Institute for Economic Research. 2007. no. 757. P. 1–30. DOI: 10.2139/ssrn.2894311.
15. Zieba M., Tomczak S.K., Tomczak J.M. Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction // Expert Systems with Applications. 2016. Vol. 58. P. 93–101. DOI: 10.1016/j.eswa.2016.04.001.
16. Xia Y., Liu C.,. Li Y, Liu N. A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring // Expert Systems With Applications. 2017. Vol. 78. P. 225–241. DOI: doi.org/10.1016/j.eswa.2017.02.017.
17. Zhou L. Performance of Corporate Bankruptcy Prediction Models on Imbalanced Dataset: The Effect of Sampling Methods // Knowledge-Based Systems. 2013. Vol. 41. P. 16–25. DOI: 10.1016/j.knosys.2012.12.007.
18. Kim T., Ahn H. A Hybrid Under-Sampling Approach for Better Bankruptcy Prediction // Journal of Intelligent Information Systems. 2015. Vol. 21, no. 2. P. 173–190. DOI: doi.org/10.13088/jiis.2015.21.2.173.
19. Veganzonesa D., Séverina E. An Investigation of Bankruptcy Prediction in Imbalanced Datasets // Decision Support Systems. 2018. no. 112. P. 111–124. DOI: 10.1016/j.dss.2018.06.011.
20. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016. P. 785–794. DOI: 10.1145/2939672.2939785.
21. Friedman J.H. Stochastic Gradient Boosting // Computational Statistics and Data Analysis. 2002. no. 38. P. 367–378. DOI: 10.1016/j.eswa.2016.04.001

Мокеев Владимир Викторович, д.т.н., старший научный сотрудник, профессор кафедры информационных технологии в экономике, Южно-Уральского государственный университет (национальный исследовательский университет) (Челябинск, Российская Федерация)

Войтецкий Роман Владимирович, студент кафедры информационных технологии в экономике, Южно-Уральского государственный университет (национальный исследовательский университет) (Челябинск, Российская Федерация)

DOI: 10.14529/cmse200305

FORECASTING ENTERPRISES BANKRUPTCY BY EXTREME GRADIENT BOOSTING

© 2020 V.V. Mokeyev, R.V. Voitetckiy
South Ural State University (pr. Lenina 76, Chelyabinsk, 454080 Russia)
E-mail: mokeyev@mail.ru, romanvoitetckii@gmail.com
Received: 27.07.2020

The application of models for forecasting bankruptcy of enterprises for controlling investment is the basis for monitoring activities of financial institutions. A crucial factor in allowing financial institutions to determine the amount of capital to cover credit losses is the accuracy of the forecast. Most studies use traditional statistical methods (for example, linear discriminant analysis and logistic regression) to build models of enterprise bankruptcy forecasting, but the accuracy of these models is usually quite low. The reason for that is the imbalanced nature of training data sets (the share of bankrupt firms is a small percent of the total number of firms). Nowadays, such machine learning methods as the random forest and the gradient boosting are becoming widespread. This study focuses on the use of extreme gradient boosting to predict bankruptcy. Extreme gradient boosting, using LASSO or Ridge regularization, penalizes complex models to avoid overfitting. Also, during training, extreme gradient boosting fills in the missing values of the data set, depending on the value of the loss. In this article, we proposed SMOTE technique to enhance the minority class of the training data sets, which helps to improve the performance of extreme gradient boosting. The experiment results of improved extreme gradient boosting are compared to the outcomes obtained by other methods.

Keywords: extreme gradient boosting, bankruptcy, enterprises, synthetic minority oversampling technique.

FOR CITATION

Mokeyev V.V., Voitetckiy R.V. Forecasting Enterprises Bankruptcy by Extreme Gradient Boosting. *Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering.* 2020. Vol. 9, no. 3. P. 77–90. (in Russian) DOI: 10.14529/cmse200305.

References

1. Altman E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance. 1968. Vol. 23, no. 4. P. 589–609. DOI: 10.1111/j.1540-6261.1968.tb00843.x.
2. Lugovskaya L. Predicting Default of Russian Smes on the Basis of Financial and Nonfinancial Variables. Journal of Financial Services Marketing. 2010. Vol. 14, no. 4. P. 301–313. DOI: 10.1057/fsm.2009.28.
3. Deakin E. A Discriminant Analysis of Predictors of Business Failure. Journal of Accounting Research. 1972. Vol. 10, no. 1. P. 167–179. DOI: 10.2307/2490225.
4. Antunesa F., Ribeiroa B., Pereirab F. Probabilistic Modeling and Visualization for Bankruptcy Prediction. Applied Soft Computing. 2017. Vol. 60. P. 831–843. DOI: 10.1016/j.asoc.2017.06.043.
5. Ohlson J.A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal Of Accounting Research. 1980. Vol. 18, no. 1. P. 109–131. DOI: 10.2307/2490395.
6. Martin D. Early Warning of Bank Failure: a Logit Regression Approach. Journal of Banking & Finance. 1977. Vol. 1, no. 3. P. 249–276.

7. Wiginton J.C. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. Journal of Financial and Quantitative Analysis. 1980. Vol. 15. P. 757–770. DOI: 10.2307/2330408.
8. Begley J., Ming J., Watts S. Bankruptcy Classification Errors in the 1980s: an Empirical Analysis of Altman's and Ohlson's Models. Review of Accounting Studies. 1996. Vol. 1, no. 4. P. 267–284. DOI: 10.1007/bf00570833.
9. Wilson R.L, Sharda R. Bankruptcy Prediction Using Neural Networks. Decision Support Systems. 1994. Vol. 11, no. 5. P. 545–557. DOI: 10.1016/0167-9236(94)90024-8.
10. Tam K.Y., Kiang M.Y. Managerial Applications of Neural Networks: the Case of Bank Failure Predictions. Management science. 1992. Vol. 38, no. 7. P. 926–947. DOI: 10.1287/mnsc.38.7.926.
11. Altman E.I., Marco G., Varetto F. Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience). Journal of Banking & Finance. 1994. Vol. 18, no. 3. P. 505–529. DOI: 10.1016/0378-4266(94)90007-8.
12. Ciampi F., Vallini C., Gordini N. Using Artificial Neural Networks Analysis for Small Enterprise Default Prediction Modeling: Statistical Evidence from Italian Firms. Oxford Business & Economics Conference Proceedings, Association for Business and Economics Research, ABER. 2009. Vol. 1. P. 126.
13. Wei L., Li J., Chen Z. Credit Risk Evaluation Using Support Vector Machine with Mixture of Kernel. Proceedings of the 7th International Conference on Computational Science. Lecture Notes in Computational Science and Engineering. 2007. Vol. 4488. P. 431–438. DOI: 10.1007/978-3-540-72586-2_62.
14. Härdle W.K., Lee Y.J., Schäfer D. The Default Risk of Firms Examined with Smooth Support Vector Machines. Discussion Papers, German Institute for Economic Research. 2007. Vol. 757. P. 1–30. DOI: 10.2139/ssrn.2894311.
15. Zieba M., Tomczak S.K., Tomczak J.M. Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications. 2016. Vol. 58. P. 93–101. DOI: 10.1016/j.eswa.2016.04.001.
16. Xia Y., Liu C.,. Li Y, Liu N. A Boosted Decision Tree Approach Using Bayesian Hyper-Parameter Optimization for Credit Scoring. Expert Systems With Applications. 2017. Vol. 78. P. 225–241. DOI: 10.1016/j.eswa.2017.02.017.
17. Zhou L. Performance of Corporate Bankruptcy Prediction Models on Imbalanced Dataset: The Effect of Sampling Methods. Knowledge-Based Systems. 2013. Vol. 41. P. 16–25. DOI: 10.1016/j.knosys.2012.12.007.
18. Kim T., Ahn H. A Hybrid Under-Sampling Approach for Better Bankruptcy Prediction. Journal of Intelligent Information Systems. 2015. Vol. 21, no. 2. P. 173–190. DOI: 10.13088/jiis.2015.21.2.173.
19. Veganzonesa D., Séverina E. An Investigation of Bankruptcy Prediction in Imbalanced Datasets. Decision Support Systems. 2018. no. 112. P. 111–124. DOI: 10.1016/j.dss.2018.06.011.
20. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016. P. 785–794. DOI: 10.1145/2939672.2939785.
21. Friedman J.H. Stochastic Gradient Boosting. Computational Statistics and Data Analysis. 2002. no. 38. P. 367–378. DOI: 10.1016/j.eswa.2016.04.001.