Сравнительный анализ моделей вероятности дефолта IT-компаний в сегменте малого и среднего бизнеса

Федотов Павел Алексеевич, студент 3-го курса финансового факультета РЭУ им. Г.В. Плеханова, г. Москва, Российская Федерация

E-mail: pavel.fedotov01.08@gmail.com

Мишин Никита Сергеевич, студент 3-го курса финансового факультета РЭУ им. Г.В. Плеханова, г. Москва, Российская Федерация

E-mail: savelgon@gmail.com

Гончаров Савелий Владимирович, студент 3-го курса финансового факультета РЭУ им. Г.В. Плеханова, г. Москва, Российская Федерация

E-mail: goncharov.saveliy@gmail.com

Аннотация

В статье рассмотрены две модели оценки кредитоспособности IT компаний малого и среднего бизнеса, функционирующих в России, – традиционная модель логистической регрессии, дающая интерпретируемый результат, и наиболее часто применимая в кредитном скоринге модель случайного леса. Результаты анализа могут быть использованы для обнаружения неблагонадежных заемщиков в кредитном портфеле.

Ключевые слова: кредитный скоринг, МСФО 9, логистическая регрессия, алгоритм случайный лес, оценка кредитного риска предприятия, IT-компании.

Comparative analysis of models of probability of default of IT-companies in the segment of small and medium business

Fedotov Pavel Alekseevich, student, Plekhanov Russian University of Economics, Moscow, Russian Federation

E-mail: pavel.fedotov01.08@gmail.com

Mishin Nikita Sergeevich, student, Plekhanov Russian University of Economics, Moscow, Russian Federation

E-mail: savelgon@gmail.com

Goncharov Savelii Vladimirovich, student, Plekhanov Russian University of Economics, Moscow, Russian Federation

E-mail: goncharov.saveliy@gmail.com

Abstract

The article contains two models for assessing the creditworthiness of IT companies of small and medium-sized businesses operating in Russia, the traditional model of logistic regression, which gives an interpretable result and the most commonly used random forest model in credit scoring. The results of the analysis can be used to detect unreliable borrowers in the loan portfolio.

Keywords: credit scoring, IFRS 9, logistic regression, random forest algorithm, assessment of the credit risk of the enterprise, IT companies.

В данной работе предполагается рассмотреть модели оценки кредитоспособности IT компаний малого и среднего бизнеса (ограничение по выручке 2 млрд рублей [4]), функционирующих в России.

В связи с вступлением в силу новых стандартов ведения банковской деятельности МСФО 9 [3] с 1 января 2018 года банковские учреждения обязаны предоставлять строгую математическую модель оценки резервов для каждого отдельно взятого займа, которую можно найти по формуле (1) [5]:

$$\mathrm{ECL} = \mathrm{PD} \times \mathrm{LGD} \times \mathrm{EAD} \quad (1)$$

где:

ECL (Expected Credit Losses) – ожидаемые потери,

PD (Probability of Default) – вероятность дефолта заемщика в будущем, вычисляемая как на основе исторических и текущих данных, так и прогнозных внешних и внутренних показателей,

LGD (Loss Given Default) – потери в случае дефолта,

EAD (Exposure at Default) – объем требований в случае наступления дефолта в будущем.

Расчет PD из предыдущей формулы является задачей бинарной классификации, которую можно решить несколькими способами. Рассматриваемый в работе метод логистической регрессии соответствует требованиям регулятора.

Разработанная в ходе написания работы модель может применяться как физическими лицами для целей P2B-кредитования, так и банками функционирующими на территории РФ для оценки кредитоспособности IT-компаний МСБ.

В выборке были оставлены лишь те компании IT-сектора, у которых присутствовали значения активов и выручки (по которым определялось наличие Баланса и ОПУ, соответственно) за все годы с 2012 до года, предшествующего дефолту (или же до 2017 года) включительно.

Данные, на которых строилась модель, содержат 8769 IT-компаний малого и среднего бизнеса, имеющих долгосрочные и краткосрочные займы, среди которых 659 компаний вышедших в дефолт, другими словами их текущий статус можно описать как «находящиеся в процедуре банкротства» или «объявлен поиск конкурсного управляющего».

Данные показатели финансовой отчетности трансформировались в длинный список факторов среди которых различные финансовые коэффициенты рентабельности, финансового рычага, ликвидности, обслуживания долга и активности, а также динамические показатели [1, 2]. Список получившихся переменных приведен в приложении.

Следующий шаг – разбиение на обучающую и тестовую выборки в соотношении 70:30 (так в обучающую выборку попало 6138, в тестовую 2631) и дальнейшая трансформация факторов с помощью WoE (Weight of Evidence) – это наиболее распространенный способ трансформации факторов[1,2], который применяется к обучающей выборке. Показатель WoE определяется следующим образом: фактор разбивается на группы (интервалы), и для каждой из сформированных групп рассчитывается значение WoE по формуле (2):

$$\mathrm{WOE_i} = \ln\left(\frac{\mathrm{N_G^{(i)}}/\mathrm{N_G}}{\mathrm{N_B^{(i)}}/\mathrm{N_B}}\right) \tag{2}$$

где:

$\mathrm{N_G^{(i)}}$ и $\mathrm{N_G}$ – количество недефолтных наблюдений в группе i и по всей выборке, соответственно,

$\mathrm{N_B^{(i)}}$ и $\mathrm{N_B}$ – количество дефолтных наблюдений в группе i и по всей выборке, соответственно.

Если значение WoE рассматриваемой группы выше нуля, это означает, что среди включенных в нее кредитов отношение недефолтов к дефолтам выше, чем во всей совокупности, и наоборот.

$$\mathrm{WOE_i} > 0 \Leftrightarrow \ln\left(\frac{\mathrm{N_G^{(i)}}/\mathrm{N_G}}{\mathrm{N_B^{(i)}}/\mathrm{N_B}}\right) > 0 \Leftrightarrow \frac{\mathrm{N_G^{(i)}}/\mathrm{N_G}}{\mathrm{N_B^{(i)}}/\mathrm{N_B}} > 1 \Leftrightarrow \frac{\mathrm{N_G^{(i)}}}{\mathrm{N_B^{(i)}}} > \frac{\mathrm{N_G}}{\mathrm{N_B}}$$

Для каждого фактора ищется такое разбиение значений на группы, чтобы разница в WoE между соседними группами была максимальной:

$$\sum|\Delta\mathrm{WOE_i}| \rightarrow \max$$

Такое разбиение наиболее достоверно отражает взаимосвязь между значениями признака и наступлением дефолта.

Пропущенные значения либо формируют отдельную группу, либо добавляются к другой группе с похожим значением WoE.

Основные преимущества WoE:

- алгоритм обработки пропущенных значений – они либо объединяются с наиболее похожей по уровню WoE группой, либо выступают в качестве отдельной группы;
- обработка экстремальных значений – экстремальные значения попадут в модель как элементы крайних групп, и им будет соответствовать значение WoE, присвоенное их группе. Иными словами, нет необходимости в отдельной, интеллектуальной обработке экстремальных значений;
- возможность обрабатывать U-образной зависимости (в отличие от линейного преобразования).

В итоге получаем следующее разбиение для параметра WoE на примере коэффициента ROA:

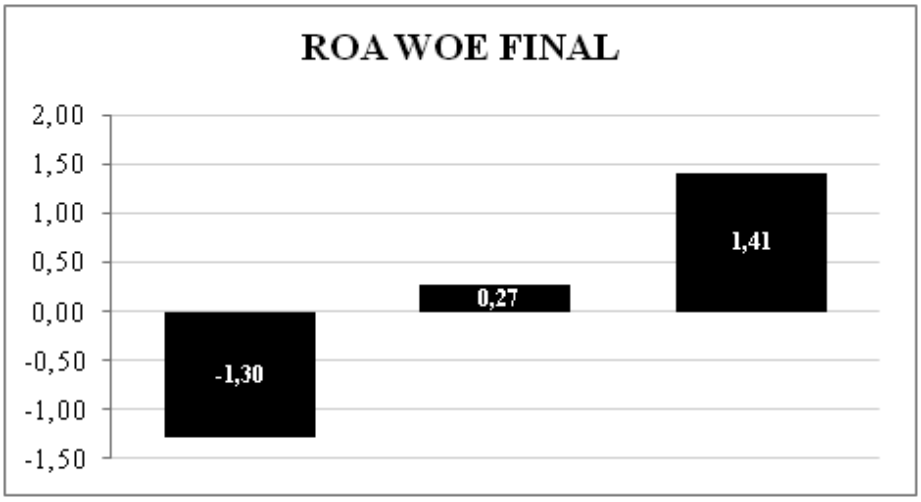

Рис. 1. Разбиение WoE

ROA < 0% → WoE = -1,30

ROA < 2,17% → WoE = 0,27

ROA ≥ 2,17% → WoE = 1,41

После разбиения всех факторов на WoE отсеиваются факторы противоречащие экономическому смыслу. Следующий шаг – оценка индивидуальной дискриминационной способности параметра с помощью показателя Джини (3):

$$\mathrm{Gini} = 2 * \mathrm{AUC} - 1, \quad (3)$$

где:

Gini – коэффициент Джини,

AUC – area under ROC curve, площадь под ROC-кривой.

И последним шагом перед построением модели является оценка парных корреляций, по следующему алгоритму – если парная корреляция больше 0,5, то из модели исключается фактор с наименьшим значением WOE. В итоге получаем короткий список факторов для

построения логистической регрессии: ROA, BDTSA, SHLR, ETL, CRPD, IC, sales_dyn, operating_profit_dyn.

Следующим шагом является построение многофакторной логистической регрессии (4):

$$PD(Y = 1|X_1, X_2, \dots X_n) = \frac{1}{1+e^{-\beta X}}, \quad (4)$$

где:

β – вектор коэффициентов логистической регрессии,

X – вектор факторов,

Y – событие дефолта заемщика в течение первого года после начала наблюдения.

В процессе отбора на каждом шаге строится регрессия по всем оставшимся на текущий момент факторам и выбрасывается показатель с наибольшим p-value, превышающим 5%, или имеющий положительный коэффициент (так как в регрессии используются WOE-факторы, которые по определению имеют отрицательную связь с уровнем дефолтов: чем выше уровень дефолтов, тем меньше WOE).

В итоге получаем следующее уравнение (5):

PD = 1 / (1 + Exp (-1 * (-4.3274172 - 0.3105799 * roawoe - 0.3912608 * bdtsawoe - 0.5204981 * crpdwoe - 0.3348378 * icwoe - 0.554516 * sdynwoe))) (5)

Таким образом, в модель вошли факторы рентабельности по чистой прибыли, отношение долга к выручке, отношение кредиторской задолженности к продажам, коэффициент покрытия процентных платежей (отношение операционной прибыли к процентным расходам) и динамика продаж.

Предсказательная способность модели на обучающей выборке составила 80,5% на тестовой 75,3% что характеризует ее как применимую для оценки резервов.

Следующим шагом стала оценка с помощью алгоритма случайного леса. Случайный лес – множество классификационных деревьев, построенных по следующему принципу:

- Каждое дерево строится по случайной подвыборке с повторениями (то есть некоторые наблюдения попадут во множество несколько раз, а некоторые не попадут вовсе)
- Дерево строится, пока создание нового узла может понизить индекс Джини дерева:

$$I_G = \sum_i \frac{n_i}{n} 2h_i(1 - h_i), \quad (6)$$

где:

n – общий размер выборки,

ni – количество наблюдений в i-ом терминальном узле,

hi - доля положительных наблюдений в i-ом терминальном узле.

Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Для решения задачи бинарной классификации были использованы соответствующие алгоритмы машинного обучения из библиотеки sklearn. Выбор параметров моделей проводился с помощью функций Grid Search (задаётся множество значений каждого параметра, после чего алгоритм с помощью перебора всех комбинаций выбирает те значения, что обеспечивают наиболее высокую точность прогнозов) и Cross Validation (использовалась пятикратная кросс-валидация). Алгоритм случайного леса после обучения имеет следующие параметры:

- оценка точности деревьев - Индекс Джини;
- при каждом ветвлении дерева случайным образом отбирается $\sqrt{k}$ штук из исходных k регрессоров;
- минимальное количество наблюдений, необходимых для разделения узла: 20;
- количество деревьев – 500.

В результате построения мы получаем коэффициент Джини равный 87%, что значительно выше, чем у модели логистической регрессии на обучающей выборке.

Таким образом, коэффициент Джини демонстрирует большую точность при оценке с помощью алгоритма случайного леса, нежели традиционной логистической регрессии и соответственно рекомендуется применять именно его для оценки кредитоспособности компании.

Список использованных источников

1. The Internal Ratings-Based Approach // Официальный сайт Bank for International Settlements [электронный ресурс] – Режим доступа. – URL: https://www.bis.org/publ/bcbsca05.pdf (дата обращения 07.07.2019).

2. Демешов Б., Тихонова А. Прогнозирование банкротства средних и малых российских компаний // Труды X Международной конференции «Применение многомерного статистического анализа в экономике и оценке качества». – ЦЕМИ РАН, 2014.

3. Едронова В., Хасянова С. Модели анализа кредитоспособности заемщиков // Финансы и кредит. – 2002. – № 6 (96).

4. Международный стандарт финансовой отчетности (IFRS) 9 «Финансовые инструменты» // Официальный сайт Министерства финансов Российской Федерации [электронный ресурс] – Режим доступа. – URL:

https://www.minfin.ru/common/upload/library/2017/02/main/MSFO_IFRS_9_1.pdf (дата обращения 07.07.2019).

5. Постановление правительства от 13 июля 2015 г. № 702 // Информационно-правовой портал ГАРАНТ.РУ [электронный ресурс] – Режим доступа – URL: http://base.garant.ru/71134484/ (дата обращения 07.07.2019).

References

1. The Internal Ratings-Based Approach // Официальный сайт Bank for International Settlements

https://www.bis.org/publ/bcbsca05.pdf

2. Demeshov B., Tikhonova A. Prognozirovanie bankrotstva srednikh i malykh rossiiskikh kompanii // Trudy X Mezhdunarodnoi konferentsii Primenenie mnogomernogo statisticheskogo analiza v ekonomike i otsenke kachestva, TsEMI RAN, 2014.

3. Edronova V., Khasyanova S. Modeli analiza kreditosposobnosti zaemshchikov. Finansy i kredit, 2002, No. 6 (96).

4. Mezhdunarodnyi standart finansovoi otchetnosti (IFRS) 9 «Finansovye instrumenty» // Ofitsial'nyi sait Ministerstva finansov Rossiiskoi Federatsii

https://www.minfin.ru/common/upload/library/2017/02/main/MSFO_IFRS_9_1.pdf

5. Postanovlenie pravitel'stva ot 13 iyulya 2015 g. № 702 // Informatsionno-pravovoi portal GARANT.RU

http://base.garant.ru/71134484/

Таблица 1

Выгружаемые факторы

Параметр	**Период (если применимо)**
Наименование	-
Регистрационный номер	-
Краткое наименование	-
Возраст компании	-
Дата ликвидации*	-
Статус	-
Регион деятельности	-
Вид деятельности/отрасль	-
Код вида деятельности (ОКВЭД)	-
Вид деятельности/отрасль (ОКВЭД КДЕС Ред. 1)	-
Код вида деятельности (ОКВЭД КДЕС Ред. 1)	-
Комментарий	-
Важная информация	-
Активы всего	2016, 2015, 2014, 2013, 2012
Амортизация*	2016, 2015, 2014, 2013, 2012
Внеоборотные активы	2016, 2015, 2014, 2013, 2012
Выручка от продажи (за минусом НДС, акцизов ...)	2016, 2015, 2014, 2013, 2012
Дебиторская задолженность	2016, 2015, 2014, 2013, 2012
Долгосрочные обязательства	2016, 2015, 2014, 2013, 2012
Запасы	2016, 2015, 2014, 2013, 2012
Капитал и резервы	2016, 2015, 2014, 2013, 2012
Кредиторская задолженность	2016, 2015, 2014, 2013, 2012
Прибыль (убыток) от продажи	2016, 2015, 2014, 2013, 2012
Проценты к уплате	2016, 2015, 2014, 2013, 2012
Чистая прибыль (убыток)	2016, 2015, 2014, 2013, 2012

Таблица 2

Аппроксимация факторов

Название переменной	**Показатель**	**Формула аппроксимации (если применимо)**
assets	Активы всего	-
non_current_assets	Внеоборотные активы	-
sales	Выручка от продажи (за минусом НДС, акцизов)	-
accounts_receivable	Дебиторская задолженность	-
long_term_liabilities	Долгосрочные обязательства	-
stocks	Запасы	-
equity	Капитал и резервы	-
accounts_payable	Кредиторская задолженность	-
operating_profit	Прибыль (убыток) от продажи	-
interests_payable	Проценты к уплате	-
net_profit	Чистая прибыль (убыток)	-
debt_short	Краткосрочный процентный долг	assets - long_term_liabilities – equity - accounts_payable
debt_long	Долгосрочный процентный долг	long_term_liabilities
debt_total	Процентный долг	debt_short + debt_long
cash_approx	Денежные средства и эквиваленты	assets - non_current_assets - stocks - accouns_receivable
capital_employed	Задействованный капитал	equity + long_term_liabilities
working_capital	Собственный оборотный капитал	equity + long_term_liabilities - non_current_assets

Таблица 3

Описание длинного списка факторов

Название переменной	Формула
NISA	net_profit / sales
ROA	net_profit / assets
ROE	net_profit / equity
RNCA	net_profit / non_current_assets
ROCE	(net_profit + interests_payable) / capital_employed
EBTA	operating_profit / assets
EBTL	operating_profit / total_liabilities
EBCE	operating_profit / capital_employed
EBSA	operating_profit / sales
BDSHE	debt_short / equity
BDLE	debt_long / equity
BDTE	debt_total / equity
BDLTA	debt_long / assets
BDTTA	debt_total / assets
TDTDE	debt_total / (debt_total + equity)
BDSHSA	debt_short / sales
BDTSA	debt_total / sales
CR	current_assets / short_term_liabilities
QR	(current_assets - stocks) / short_term_liabilities
CASH_R	cash_approx / short_term_liabilities
CATL	current_assets / total_liabilities
CAS	current_assets / sales
CLTA	short_term_liabilities / assets
NCATA	non_current_assets / assets
SHLR	equity / long_term_liabilities
SR	equity / assets
WCTA	working_capital / assets
CS	cash_approx / sales
LIQ	cash_approx / assets
CDTA	(cash_approx + accounts_receivable) / assets
NWCCA	working_capital / current_assets
ETL	equity / total_liabilities
CLCTA	(current_liabilities - cash_approx) / assets
LTLTA	long_term_liabilities / assets
CRPD	accounts_payable / sales

Название переменной	**Формула**
CLPD	accounts_receivable / sales
IC	operating_profit / interests_payable
SNCA	sales / non_current_assets
NAS	sales / capital_employed
SS	sales / stocks
SATA	sales / assets
assets_dyn	(assets текущего года – assets предыдущего года) / assets предыдущего года
sales_dyn	(sales текущего года – sales предыдущего года) / sales предыдущего года
equity_dyn	(equity текущего года – equity предыдущего года) / equity предыдущего года
operating_profit_dyn	(operating_profit текущего года – operating_profit предыдущего года) / operating_profit предыдущего года
net_profit_dyn	(net_profit текущего года – net_profit предыдущего года) / net_profit предыдущего года