УДК 65.012.8 JEL C55 DOI 10.26425/1816-4277-2019-9-48-53

Баклушинский Вадим Валентинович
соискатель, ФГБОУ ВО «Ульяновский государственный университет», г. Ульяновск, Российская Федерация
e-mail: *vbaklushinskiy@mail.ru*

Пустынникова Екатерина Васильевна
д-р экон. наук, ФГБОУ ВО «Ульяновский государственный университет», г. Ульяновск, Российская Федерация
e-mail: *ebrezneva@list.ru*

МАШИННОЕ ОБУЧЕНИЕ КАК ИНСТРУМЕНТ КОРПОРАЦИИ ДЛЯ ВЫБОРА ПОСТАВЩИКОВ

Аннотация. *В области экономики и финансов, методы машинного обучения получили распространение при решении проблем исследования поведения потребителей и в торговле валютой и ценными бумагами. Тем не менее, они слабо развиты в решении вопросов, связанных со взаимодействием между предприятиями. В статье представлены результаты составления и тестирования моделей машинного обучения, созданных в целях оценки благонадежности предприятий как поставщиков. Исходя из проведенного анализа, методы машинного обучения применимы при проведении оценки поставщиков. Эта статья написана на тему расширения области применения машинного обучения в сфере анализа поведения коммерческих предприятий.*

Ключевые слова: *машинное обучение, оценка поставщиков, большие данные, классификация, экономическая безопасность, управление корпорацией.*

Цитирование: Баклушинский В.В., Пустынникова Е.В. Машинное обучение как инструмент корпорации для выбора поставщиков//Вестник университета. 2019. № 9. С. 48-53.

Baklushinskii Vadim
Postgraduate student, Ulianovsk State University, Ulianovsk, Russia
e-mail: *vbaklushinskiy@mail.ru*

Pustynnikova Ekaterina
Doctor of Economic Sciences, Ulianovsk State University, Ulianovsk, Russia
e-mail: *ebrezneva@list.ru*

MACHINE LEARNING AS A CORPORATION'S TOOL FOR SELECTION OF SUPPLIERS

Abstract. *In the economics and finance, machine learning methods have spread when solving the problems of consumer behavior research and in currency and securities trading. However, they are poorly developed in dealing with issues related to interaction between enterprises. The article presents the results of the compilation and testing of machine learning models, created to assess the reliability of enterprises as suppliers. According to the analysis, carried out in the article, machine learning methods are applicable when conducting supplier evaluations. This article has been written on the theme of expanding the scope of machine learning in the field of analysis of the behavior of commercial enterprises..*

Keywords: *machine learning, suppliers' assessment, big data, classification, economic security, corporate management.*

For citation: Baklushinskii V.V., Pustynnikova E.V. Machine learning as a corporation's tool for selection of suppliers (2019) Vestnik universiteta, I. 9, pp. 48-53. doi: 10.26425/1816-4277-2019-9-48-53

Методы машинного обучения появились в 1950 гг., впервые были описаны американскими инженерами А. Сэмюэлем, Дж. Вейценбаумом и ученым Ф. Розенблаттом и вплоть до начала 2010 гг. не находили массового применения в развитии каких-либо отраслей экономики. Столь большой временной разрыв между созданием методов и началом их массового использования в производстве услуг может быть объяснен двумя причинами. В первую очередь это техническая невозможность реализации методов для решения сложных задач с теми ограниченными мощностями, которыми обладали компьютеры во второй половине XX в. Вторая причина заключается в отсутствии до 2010 гг. необходимого объема больших данных для обучения программ.

В начале XXI в. произошел существенный рост количества пользователей сети «Интернет», что многократно увеличило объемы доступной информации о потенциальных покупателях товаров, потребительских предпочтениях и т. п. и стимулировало применение методов машинного обучения в анализе рынка. Помимо данных о потребительском поведении в сети увеличился и объем информации о предприятиях. Эти данные

могут позволить провести обучение программ, способных прогнозировать поведение компаний. В частности, таким способом можно решать проблему оценки и выбора поставщиков.

Выбор поставщиков определяет не только издержки промышленного предприятия, но и бесперебойность производства, и его экономическую безопасность. Обычно этот процесс требует вынесения экспертных оценок специалистами нескольких подразделений предприятия, что несет с собой риски, связанные с влиянием человеческого фактора.

Если компании, которые расценивают на роль поставщика, действуют достаточно давно, и у них сложилась определенная репутация, данный подход к выбору вполне оправдан. Проверка благонадежности предприятия как поставщика может быть проведена исходя из истории его взаимоотношений с другими организациями: подавались ли в адрес организации судебные иски по поводу неисполнения обязательств по договору, находится ли она в перечне недобросовестных поставщиков, определенном Федеральной антимонопольной службой России (далее – ФАС России), достоверна ли ее контактная информация и находится ли она в состоянии банкротства.

Также может быть изучена репутация управляющих предприятия: есть ли они в списке дисквалифицированных лиц, составленном Федеральной налоговой службой России (далее – ФНС России), привлекались ли они к ответственности за нарушение законодательства и т. п. Применима для проверки благонадежности и информация о величине уставного капитала предприятия, платежеспособности и его производственных показателях.

Перечисленные источники информации о предприятии можно разделить на исторические данные (вся информация, которая касается истории предприятия, его регистрационных данных и репутации его руководства) и его текущие данные (информация относительно текущего состояния предприятия). Если предприятие существует достаточно долго, оно оставляет в открытых источниках большие объемы информации, которая может быть использована для комплексной оценки его благонадежности.

Однако в случаях, когда предприятие работает непродолжительный промежуток времени, вся доступная информация о нем сводится к данным, опубликованным на сайтах государственных служб и той информации, которую само предприятие готово передать своим контрагентам. И именно в такой ситуации, когда оценка благонадежности стандартными методами не позволяет сделать однозначных выводов, могут быть применены методы машинного обучения.

Использование методов машинного обучения в аналитике по сфере b2b стало объектом исследования ряда ученых. В частности, M. Bohanec, M Robnik-Sikonja, M. Kljajic Borstnar описали и применили модель машинного обучения, которая прогнозирует объем продаж продуктов компании в сторону юридических лиц [7]. K. Stormi, T. Laine, T. Elomaa в своей статье показали возможность применения методов машинного обучения, которые широко используются в анализе потребительского рынка, при исследовании продаж юридическим лицам [11]. Перечисленными выше авторами были разработаны модели, способные сегментировать рынок и прогнозировать список активных клиентов в сфере b2b.

Y. Karlinsky Shichor опубликовала результаты применения модели машинного обучения в ценообразовании на продукцию компании, выполняющей оптовые поставки алюминия [10]. Используя гибридный подход, сочетающий принятие решений непосредственно продавцом при нетипичных сделках и ценообразование с помощью разработанной модели в стандартных случаях, указанный выше автор объявила о существенном приросте маржи от сделок с предприятиями.

Научная новизна исследования состоит в доказательстве возможности и целесообразности использования методов машинного обучения в оценке благонадежности контрагентов.

Цель исследования – определить и интерпретировать долю верных оценок благонадежности поставщиков, данных при помощи различных методов машинного обучения. Для достижения цели были поставлены следующие задачи: обучить и протестировать модели оценки поставщиков, составленные по различным алгоритмам; проверить достоверность результатов, полученных при тестировании моделей; сделать выводы о целесообразности применения различных видов моделей в решении задачи оценки благонадежности поставщиков.

Теоретическая значимость работы состоит в развитии темы применения методов машинного обучения к решению вопросов взаимодействия между предприятиями. Практическая значимость исследования заключается в возможности использования примененной авторами методики исследования для оценки поставщиков корпорациями.

По своей сути, оценка благонадежности предприятия является вопросом классификации его как организации, не склонной к нарушению договорных обязательств или же как неблагонадежной организации. Для решения задач классификации методами машинного обучения используются классификаторы, а именно алгоритмы логистической регрессии (англ. logistic regression), метод k-ближайших соседей (англ. k-nearest neighbors) и метод дерева решений (англ. decision trees) [4]. Классификатор обучается на наборе данных (атрибутах и присвоенных им классах), после чего может присваивать классы новым объектам, у которых их еще нет. Приведем описание этих алгоритмов.

Логистическая регрессия определяет принадлежность объекта к одному из двух классов исходя из вычисленных по его числовым характеристикам значений функции логистической кривой. Логистическая кривая представляет собой график возрастающей нелинейной функции, имеющий форму, схожую с символом «S». Значения этой функции определены на промежутке от 0 до 1, и, округленные до целого числа, отражают принадлежность объекта к одному из двух классов (в контексте этой статьи, класс «0» – благонадежный контрагент, «1» – неблагонадежный). Обучение модели производится при помощи корректировки коэффициентов независимых переменных под значения функции, близкие к присвоенным объектам классам, методом градиентного спуска [1; 9].

Алгоритм k-ближайших соседей состоит в присвоении новому объекту того же класса, что и у большинства других объектов, близких к нему по своим числовым характеристикам. Для этого, рассчитывается расстояние между значениями независимых переменных нового объекта и объектов, уже имеющих класс. Расстояние между объектами определяется по формуле Евклидова расстояния, а именно, как квадратный корень из суммы квадратов разности координат объектов.

Дерево решений – по данному алгоритму выстраивается классификатор, состоящий из листьев (непосредственно классов), которые присваивает модель, и узлов проверки (критериев, по которым классифицируются объекты). Листья и узлы проверки располагаются в иерархической структуре, которая позволяет классифицировать по себе новые объекты. Процесс составления дерева решений заключается в группировке объектов по значениям их атрибутов.

Одним из вариантов этого метода является случайный лес [2; 12]. Его отличие состоит в том, что из набора данных, используемых для тренировки алгоритма, случайным образом выбираются несколько наборов меньшего размера, по которым строятся разные деревья. Класс объекта определяется как усредненный ответ построенных деревьев. Таким образом, устраняется часть ошибок, которые производятся единичными деревьями.

Все приведенные выше методики для своей реализации требуют получения вводных данных по большому количеству объектов. В случае сформулированной нами задачи, для обучения модели необходима историческая информация по значительному количеству предприятий. При этом необходим как набор данных, соответствующий заранее известным характеристикам предприятий (среднесписочная численность, обороты, данные по налоговой задолженности, величина уставного капитала и т. д.), так и информация об их поведении в исторической перспективе (были ли случаи неисполнения предприятиями договорных обязательств). После обучения модели, она будет способна с той или иной точностью классифицировать предприятие по доступным его контрагентам данным.

Информация подобного рода, как правило, не публикуется предприятиями, если того не требует законодательство, а в ряде случаев (например, данные о величине задолженности по налогам или об эпизодах нарушения договоров) ими скрывается. По этой причине, читатель статьи может счесть, что информация, необходимая для обучения модели оценки благонадежности, не может быть доступна к использованию. Но напротив, большие объемы информации о деятельности предприятий уже находятся в открытом доступе и регулярно обновляются государственными службами России. В частности, на сайте ФАС России опубликован перечень неблагонадежных поставщиков [5]. ФНС России публикует в электронном виде данные о среднесписочной численности предприятий, их оборотах и расходах за предыдущие годы, величине долга по налогам и сборам, а также список дисквалифицированных лиц [6].

В случае реализации методик машинного обучения заинтересованными в оценке своих поставщиков предприятиями, ими могут использоваться не только данные, находящиеся в открытом доступе, но и собственные базы по контрагентам. Например, достаточно долго существующее предприятие может обладать сведениями о величине уставного капитала, остатках денежных средств по финансовой отчетности и т. п. большого числа своих контрагентов. А эта информация, в настоящий момент малодоступная для третьих лиц, может позволить создавать наиболее точные модели машинного обучения.

Для подтверждения эффективности предложенной оценки благонадежности поставщика методами машинного обучения, авторами было проведено исследование на наборе больших данных. Источником данных для исследования стала открытая информация, опубликованная на сайтах ФАС России и ФНС России.

Исследование состояло из следующих этапов.

1. Поиск и извлечение из открытых источников выборки для обучения моделей. В качестве атрибутов были приняты данные ФНС России о среднесписочной численности работников предприятий, задолженности предприятий по налогам и сборам, а также валовой рентабельности предприятий за год (соответственно, наборы данных № 75, № 79 и № 77 с сайта ФНС России) [6]. В качестве ярлыков для классификации использованы данные из перечня неблагонадежных поставщиков, которые нарушили условия договоров при исполнении государственного заказа. Этот перечень опубликован на сайте ФАС России [5].

2. Обработка данных для их приведения в форму, пригодную для их использования в моделях. Перечисленные выше наборы данных были опубликованы в виде заархивированных файлов, для объединения которых в исходные таблицы использовано программное обеспечение (далее – ПО) Microsoft Office Excel, включая пакет разработчика VBA. Соединение данных из различных таблиц в единый набор проведено при помощи ПО Anaconda.

3. Формирование обучающей выборки из полученного на предыдущем этапе набора. В качестве выборки из таблицы извлечены данные по 730 предприятиям, нарушившим условия контрактов с государством, а также случайно выбранным 730 предприятиям, которые отсутствовали в перечне неблагонадежных поставщиков от ФАС России.

4. Выборка была разделена на обучающую и тестовую выборки, с соотношением числа объектов соответственно 80 и 20 %. Тестовая выборка использовалась на следующем шаге для проверки точности классификации, сделанной моделями.

5. Были обучены модели логистической регрессии, k-ближайших соседей, дерева решений и случайного леса. После обучения точность классификаторов была проверена на тестовой выборке предприятий как отношение числа случаев верных предсказаний модели к общему размеру тестовой выборки. Для создания классификаторов было использовано ПО Anaconda (библиотека «scikit-learn») [3].

6. Выдвинута нулевая гипотеза (H_0) об отсутствии статистически значимой связи между предсказанными при помощи моделей и фактически присвоенными классами. Таким образом, делается предположение о том, что присвоение моделью класса объектам выполняется случайно.

По формуле полной вероятности математическое ожидание случайного правильного выбора класса при вероятности принадлежности объекта к одному из двух классов равной 0,5 и вероятности *x* для выбора классификатором этого класса равно 0,5, расчет дан в формуле (1):

$$p = 0{,}5 \cdot x + (1 - 0{,}5) \cdot (1 - x) = 0{,}5. \quad (1)$$

Также выдвинута альтернативная гипотеза (H_1) – между предсказанными моделью и фактическими классами есть значимая связь, и точность предсказаний модели выше 50 %. Поскольку две из этих моделей показали долю верных предсказаний выше 60 %, была сформулирована еще одна гипотеза (H_2), заключающаяся в том, что между предсказанными и фактически присвоенными классами есть связь и точность предсказаний модели выше 60 %.

Нулевая гипотеза проверена при помощи биномиального теста средствами ПО Anaconda [8]. Результаты тестирования моделей приведены в таблице 1.

Таблица 1

Точность предсказаний классификаторов в ходе исследования

Алгоритм, по которому действовал классификатор	**Случаи верных предсказаний модели, %**	**Вероятность H_0 по биномиальному тесту, с альтернативной гипотезой H_1, %**	**Вероятность H_0 по биномиальному тесту, с альтернативной гипотезой H_2, %**
Логистическая регрессия	49,8	52,3	99,9
k-ближайших соседей (n=2)	61,0	0,0	39,3

Окончание табл. 1

Алгоритм, по которому действовал классификатор	Случаи верных предсказаний модели, %	Вероятность H_0 по биномиальному тесту, с альтернативной гипотезой H_1, %	Вероятность H_0 по биномиальному тесту, с альтернативной гипотезой H_2, %
Дерево решений	58,5	0,2	71,4
Случайный лес (n=3)	65,2	0,0	4,3

Составлено авторами по материалам исследования

Классификатор на основе логистической регрессии дал тривиальный результат, поскольку нулевая гипотеза подтвердилась с вероятностью в 0,523. Тест нулевой гипотезы по остальным классификаторам (k-ближайших соседей, дерево решений и случайный лес) показал ее вероятность менее 5 %, что говорит о наличии статистически значимой связи между предсказаниями классификатора и классами объектов и возможности применения этих классификаторов для оценки благонадежности поставщика. Наибольшую эффективность показал алгоритм случайного леса, включившего в себя три дерева: его результат лучше результатов модели одиночного дерева решений на 6,7 % по причине исключения части ошибок за счет усреднения результатов. Алгоритм ближайших соседей также оказался достаточно эффективным в решении поставленной задачи.

Поскольку доля верных предсказаний по двум классификаторам выше 60 %, был проведен дополнительный биномиальный тест с предположением, что точность классификаторов более 60 %. Данное предположение оказалось верным только в случае классификатора с алгоритмом случайного леса при уровне значимости в 5 %. Из этого следует вывод, что при обучении на использованном авторами наборе данных только классификатор, использующий алгоритм случайного леса, работает с достоверной точностью более 60 %.

Тем не менее, точность присвоения предприятию верного класса в 60 % случаев недостаточна для однозначного вывода о том, что рассматриваемое в качестве поставщика предприятие является неблагонадежным поставщиком. Для повышения точности работы, классификатор должен быть обучен на релевантных данных из других источников, в том числе, на собственных базах данных заинтересованных в развитии таких классификаторов предприятий.

Использование методов машинного обучения для анализа поведения поставщиков, предложенное в этой статье, может применяться промышленными предприятиями. Уже на текущий момент в открытом доступе находится достаточно большой объем информации о предприятиях, который может быть использован для создания классификаторов, оценивающих поведение контрагентов. Проведенное авторами исследование подтвердило возможность использования машинного обучения для этих целей. Наиболее эффективным алгоритмом машинного обучения в целях решения задачи оценки благонадежности поставщика является случайный лес, по которому был создан классификатор, работающий с точностью более 60 %.

Учитывая тренд к увеличению доступности информации о деятельности организаций, в чем немаловажную роль играют российские государственные органы, авторы делают вывод о том, что машинное обучение будет широко использоваться для анализа поведения предприятий уже в ближайшем будущем.

Библиографический список

1. Алексеева, В. А. Использование методов машинного обучения в задачах бинарной классификации//Автоматизация процессов управления. – 2015. – № 3 (41). – С. 58-63.
2. Картиев, С. Б. Алгоритм классификации, основанный на принципах случайного леса, для решения задачи прогнозирования / С. Б. Картиев, В. М. Курейчик//Программные продукты и системы. – 2016. – № 2. – С. 11-15.
3. Коротеев, М. В. Обзор некоторых современных тенденций в технологии машинного обучения//E-Management. – 2018. – № 1. – С. 26-35.
4. Краснянский, М. Н. и др. Сравнительный анализ методов машинного обучения для решения задачи классификации документов научно-образовательного учреждения / М. Н. Краснянский, А. Д. Обухов, Е. М. Соломатина, А. А. Воякина// Вестник Воронежского государственного университета. – 2018. – № 3. – С. 173-182.

5. Информационные системы//Федеральная антимонопольная служба [Электронный ресурс]. – Режим доступа: https://fas.gov.ru/pages/about/about/gositsystem.html (дата обращения: 27.06.2019).
6. Открытые данные//Федеральная антимонопольная служба [Электронный ресурс]. – Режим доступа: https://www.nalog.ru/rn77/opendata/ (дата обращения: 27.06.2019).
7. Bohanec, M. [et al.]. Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting / M. Bohanec, M. Robnik-Sikonja, M. Kljajic Borstnar//Organizatsiya. – 2017. – № 50 (3). – Pp. 217-233.
8. Breheny, P. [et al.] p-Value Histograms: Inference and Diagnostics / P. Breheny, A. Stromberg, J. Lambert//High-Throughput. – 2018. – Vol. 7. – № 3 [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/publication/327356617_p-Value_Histograms_Inference_and_Diagnostics (дата обращения: 27.06.2019).
9. Farrelly, C. M. Topology and Geometry in Machine Learning for Logistic Regression//PsyArXiv. – Oct. 17, 2017 [Электронный ресурс]. – Режим доступа: psyarxiv.com/v8jgk (дата обращения: 27.06.2019).
10. Karlinsky Shichor, Y. Automation, Decision Making and Business to Business Pricing//Columbia University. – July 1, 2018 [Электронный ресурс]. – Режим доступа: https://academiccommons.columbia.edu/doi/10.7916/D83X9Q5M (дата обращения: 27.06.2019).
11. Stormi, K. [et al.]. Feasibility of b2c customer relationship analytics in the b2b industrial context / K. Stormi, T. Laine, T. Elomaa// Research papers. – 2018. – № 61. – Pp. 1-8.
12. Zieba, M. [et al.]. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction / M. Zieba, S. K. Tomczak, J. M. Tomczak//Expert Systems with Applications. – 2016. – Vol. 58. – Pp. 93-101.

References

1. Alekseeva V. A. Ispol'zovanie metodov mashinnogo obucheniya v zadachakh binarnoi klassifikatsii [*The use of machine learning methods in binary classification*]. Avtomatizatsiya protsessov [*Automation of management processes*], 2015, I. 3 (41), pp. 58-63.
2. Kartiev S. B., Kureichik V. M. Algoritm klassifikatsii, osnovannyi na printsipakh sluchainogo lesa, dlya resheniya zadachi prognozirovaniya [*Classification Algorithm based on the Principles of a Random Forest, for Solving the Prediction Problem*]. Programmnye produkty i sistemy [*Software Products and Systems*], 2016, I. 2 (114), pp. 11-15. Available at: https://cyberleninka.ru/article/n/algoritm-klassifikatsii-osnovannyy-na-printsipah-sluchaynogo-lesa-dlya-resheniya-zadachi-prognozirovaniya (accessed 27.06.2019).
3. Koroteev M. V. Obzor nekotorykh sovremennykh tendentsii v tekhnologii mashinnogo obucheniya [*Review of some contemporary trends in machine learning technology*]. E-Management, 2018, I. 1, pp. 26-35.
4. Krasnyanskii M.N., Obukhov A.D., Solomatina E.M., Voyakina A.A. Sravnitel'nyi analiz metodov mashinnogo obucheniya dlya resheniya zadachi klassifikatsii dokumentov nauchno-obrazovatel'nogo uchrezhdeniya [*Comparative analysis of machine learning methods for solving the problem of classification of documents of a scientific and educational institution*]. Vestnik Voronezhskogo gosudarstvennogo universiteta [*Bulletin of Voronezh State University*], 2018, I. 3, pp. 173-182. Available at: http://www.vestnik.vsu.ru/pdf/ analiz/2018/03/2018-03-19.pdf (accessed 27.06.2019).
5. Informatsionnye sistemy [*Data systems*]. Federal'naya antimonopolnaya sluzhba [*Federal antimonopoly service*]. Available at: https://fas.gov.ru/pages/about/about/gositsystem.html (accessed 27.06.2019).
6. Otkrytye dannye [*Open data*]. Federal'naya nalogovaya sluzhba [*Federal tax service*]. Available at: https://www.nalog.ru/rn77/opendata/ (accessed 27.06.2019).
7. Bohanec M., Robnik-Sikonja M., Kljajic Borstnar M. Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting. Organizatsiia, 2017, I. 50 (3), pp. 217-233.
8. Breheny P., Stromberg A., Lambert J. p-Value Histograms: Inference and Diagnostics. High-Throughput, 2018, Vol. 7, I. 3. Available at: https://www.researchgate.net/publication/327356617_p-Value_Histograms_Inference_and_Diagnostics (accessed 27.06.2019).
9. Farrelly C. M. Topology and Geometry in Machine Learning for Logistic Regression. PsyArXiv. Oct. 17, 2017. Available at: psyarxiv.com/v8jgk (accessed 27.06.2019).
10. Karlinsky Shichor Y. Automation, Decision Making and Business to Business Pricing. Columbia University. July 1, 2018. Available at: https://academiccommons.columbia.edu/doi/10.7916/D8058ZDR/download (accessed 27.06.2019).
11. Stormi, K., Laine, T., Elomaa, T. Feasibility of b2c customer relationship analytics in the b2b industrial context. Research papers, 2018, I. 61, pp. 1-8.
12. 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, pp. 93-101.