Доц. д-р Антон Герунов*

ПРИЛОЖЕНИЕ НА КЛАСИФИКАЦИОННИ АЛГОРИТМИ ЗА МОДЕЛИРАНЕ НА ИКОНОМИЧЕСКИ ИЗБОРИ

Изследвана е възможността за прилагане на алгоритми от сферата на машинното самообучение за решаване на икономически задачи с дискретен двоичен избор. За целта са разгледани три типични задачи, стоящи пред бизнеса: класификация при операции с отпускане на овърдрафт по кредитни карти, при управление на кредитния риск и при провеждане на маркетингови кампании. Тези ситуации са моделирани както с класически статистически методи (логистична регресия и линеен дискриминантен анализ), така и чрез пет метода за машинно самообучение (невронни мрежи, kNN-алгоритъм, наивен бейсов класификатор, случайна гора и машина с подкрепящи вектори). Основният резултат е, че при всяка от разгледаните задачи моделът на случайна гора устойчиво регистрира по-висока прогностична точност спрямо обичайно използваните иконометрични методи. Този извод обосновава нуждата от разширение на иконометричния инструментариум чрез интегриране на нови класификационни методи, като случайната гора, машината с подкрепящи вектори и невронната мрежа се очертават като водещи алтернативи.[1]

JEL: C45; C53; D81

Ключови думи: класификация; двоичен изход; икономически избор; алгоритми за машинно самообучение; класификационна точност

Икономическите задачи, свързани с избор или класификация, са често срещани както в научните изследвания, така и в стопанската практика (Hensher & Johnson, 2018). Сред тях се откроява проблемът на потребителския избор между дискретни алтернативи, прогнозиране на потребителско поведение, управление на кредитни и ликвидни рискове, класификация на физически и юридически лица по дадена характеристика и много други. Стандартният иконометричен инструментариум за моделиране на подобни задачи включва инструменти като логистичната регресия и линейния дискриминантен анализ. Макар че ползата

* СУ „Кл. Охридски“, Стопански факултет, A.Gerunov@feb.uni-sofia.bg

[1] Assoc. Prof. Anton Gerunov, Ph.D. CLASSIFICATION ALGORITHMS FOR MODELING ECONOMIC CHOICE. *Summary*: The article shows how some novel machine learning algorithms can be applied to economic problems of discrete binary choice. An examination is made of three typical business tasks – classifying overdraft applications, credit risk management, and marketing segmentation. Both traditional econometric methods (logistic regression and linear discriminant analysis) as well as five more advanced machine learning algorithms (neural networks, k-nearest neighbours, naïve Bayes classifier, random forest, and support vector machine) have been used for modelling these tasks. For all the classification tasks, the random forest algorithm robustly registers improved forecasting accuracy over the more traditional approaches. This underlines the need to supplement the classical econometric toolbox with innovative methods, with the random forest, the support vector machine, and the neural network being prime candidates. *Keywords: classification; binary choice; economic choice; machine learning methods; classification accuracy*.

от тях е значителна, те срещат известни затруднения при моделирането на особено големи масиви от данни, които са типични за съвременните икономически дейности. Паралелно с това се полагат активни изследователски усилия за извеждането и апробирането на нови методи в сферата на машинното самообучение, които биха били приложими и за моделирането на икономически задачи с класификационен характер (Hastie et al., 2005).

Целта на изследването е да се сравнят класическите подходи за подобно моделиране на дискретен избор с пет метода за машинно самообучение – невронна мрежа, дървета на решенията, обединени в случайна гора, алгоритъм k-най-близки съседи, бейсова класификация и машина с подкрепящи вектори. Те са анализирани в контекста на три икономически ситуации – управление на риска от неплащане при операции с кредитни карти, управление на кредитния риск и при провеждане на директна маркетингова кампания. За сравняване на различните алгоритми е използвана тяхната относителна прогностична точност, като се следва допускането, че алгоритмите с по-висока прогностична точност са и по-подходящи за моделиране на дадената рискова ситуация.

Класически методи за моделиране на двоичен избор

Решаването на класификационни задачи има продължителна история, като в икономиката води началото си от моделирането на стопанските избори между краен брой алтернативи. Един от първите методи, свързани с това, е логистичната регресия, която използва логистична функция, за да определи принадлежността към определен клас. Сред ранните приложения на метода заслужава да се отбележат работите на Cox (1958) и McFadden (1981), в които ключовият въпрос е да се моделира вероятността (P) дадено наблюдение (y) да принадлежи към определен клас, ако са известни и други характеристики на това наблюдение (x_i). Тази условна вероятност е означена като $P(y|x_i)$, а вероятността наблюдението да е от даден клас се изчислява с помощта на логистичната функция:

$$(1) \qquad P(y|x_i) = \frac{\exp(\beta_0 + \Sigma\beta_i x_i)}{1 + \exp(\beta_0 + \Sigma\beta_i x_i)}$$

Силата на връзката между зависимата и независимите променливи е отразена в размера на изчислените бета коефициенти в уравнение (1), като интерпретацията е следната: с увеличаването на размера на тези коефициенти се повишава и вероятността за класификация в нарастващ клас спрямо избрания базисен такъв. Макар първоначално логистичната регресия да е използвана главно при случаите на дискретна зависима променлива с две потенциална стойности, тя може да бъде разширена и до многомерна логистична регресия, решаваща класификационната задача между няколко типа класове[2].

[2] За повече детайли и приложения вж. Hyman & Yang (2001) и Akinci et al. (2007).

Друг класически подход за решаване на класификационни задачи, който работи отново с дискретни променливи, е линейният дискриминантен анализ. Както предполага и името му, това е линеен метод, който цели да раздели класификационната равнина на две или повече подравнини, като всяка от тях съдържа представителите на един определен клас. По същество методът конструира оптималната линейна комбинация от обяснителни променливи, която може да бъде използвана за определяне на принадлежност към даден клас.

По-нататък ще представим накратко този алгоритъм с една зависима и една обяснителна променлива.[3] Отново обозначаваме зависимата променлива с *y*, а независимата с *x*. Обозначаваме двете условни вероятностни разпределения с *p(y|x)* за вероятността да наблюдаваме *y,* ако наблюдаваме *x,* и съответно – с *p(x|y)* за вероятността да наблюдаваме *x,* ако наблюдаваме *y*. Допускаме, че двете променливи следват нормалното разпределение, като обозначаваме средните стойности съответно с μ_y и μ_x, а техните ковариации съответно с σ_{yx} и σ_{xy}. При допускането, че с σ_{yx} и σ_{xy} означаваме съответните ковариации на двете променливи, можем да осъществим класификация на базата на следното условие (с *T* е обозначена определена прагова стойност):

$$(2) \qquad \left(x-\mu_y\right)^T \sigma_{yx}^{-1}\left(x-\mu_y\right) - (x-\mu_x)^T \sigma_{xy}^{-1}(x-\mu_x) < T.$$

Поради това, че този метод е сравнително интуитивен за приложение и тълкуване, той продължава да се използва в някои отделни случаи. Извън регресионните модели и линейните дискриминантни модели тук съществуват и множество други традиционни статистически методи, които могат да бъдат използвани за управление на рисковете. Редица от тях са описани в класическата работа на Hastie et al. (2005; 2013), където те са сравнени и с по-съвременните алгоритми от сферата на машинното самообучение. Прави впечатление, че традиционните методи все още се използват, макар в общия случай да имат по-ниска прогностична точност, отколкото по-съвременните подходи. Вероятно това се дължи в значителна степен на зависимостта от пътя на развитие на аналитичните дисциплини – тези методи са познати, добре проучени, широко прилагани и много експерти имат опит и умения предимно с тях.

Алгоритми за машинно самообучение за моделиране на двоичен избор

Алгоритмите за надзиравано машинно самообучение се характеризират с това, че те имат нужда от маркирани данни с ясно разграничени класове (или стойности) на целевата променлива. Това най-често включва предварителна обработка на данните (от хора или машинна), която да определи дали целевата

[3] За повече детайли и пълно представяне на алгоритъма вж. Ripley (1996).

променлива принадлежи на даден клас преди етапа на самото моделиране. Макар най-простото маркиране да е двоична (бинарна) променлива, няма причина етикетите да не са с по-голям брой значения, за да се отчетат нюансите на реализациите. Най-често използваните в литературата и в практиката алгоритми за надзиравано обучение са невронни мрежи, k-най-близки съседи, бейсови подходи, дървета за взимане на решения, обединени в случайни гори, и машини с подкрепящи вектори, но се прилагат също и традиционни статистически подходи като логистичната регресия и дискриминантния анализ (Chandola et al., 2009; Phua et al., 2010; Omar et al., 2013; Qiu et al., 2016; Rousseeuw & Hubert, 2018). Важно е да се подчертае, че навлизат и много нови потенциално полезни алгоритми, като доста от тях са различни варианти на вече изброените.

Невронни мрежи

Невронните мрежи са изчислителни алгоритми, чиято структура е силно повлияна от начина, по който функционира човешкият мозък. В статистическата невронна мрежа архитектурата на алгоритъма е сходна с тази на мозъка, като ролята на неврони играят различни променливи и стойности, а активацията се извършва посредством предварително зададена математическа функция, чрез която се правят изчисления и се предават различните стойности в рамките на модела. Целта на подобен модел е въз основа на група от зависими променливи (поредица x_1, x_2, …, x_n) да прогнозира стойността или класа на целева независима променлива (обозначена с y). Обяснителните зависими променливи формират входния слой на невронната мрежа.

Всяка от тези променливи влияе върху оценката на крайната целева променлива чрез поредица от претеглени стойности на базата на предварително определени функции. Накратко, входният слой предава активиращи импулси, изчислени според определена функция на активация K, към първия междинен слой. Той от своя страна използва тези импулси като входни данни за функциите си на активация към следващия слой и така до последния, който определя крайната целева променлива. В този смисъл невронната мрежа може да се представи чрез нейната мрежова функция f. Ако обозначим даден входен неврон (обяснителна променлива) с x, а функцията на активация с K, то мрежовата функция на невронната мрежа може да бъде представена по следния начин:

$$(3) \qquad f(x) = K\left(\sum_{i=1}^{n} w_i\, g_i(x)\right).$$

Тук с w_i са обозначени поредицата от тегла на импулсите, изпратени от тази променливи към междинните и крайния слой, а с $g_i(x)$ – поредица от математически функции, претеглени спрямо теглата w_i. K е предварително определена функция на активация (често се използва сигмоидна функция). На практика оценката на модел (обучението) на невронна мрежа се състои от изчисляване на теглата при дадени ограничения – използваните математически функции и желания брой междинни слоеве. Това означава, че при един и същи

набор от данни могат да бъдат обучени множество различни невронни мрежи, всяка от които с различни характеристики. Изборът на оптималната мрежа се определя от редица фактори, но трябва да се акцентира най-вече върху възможността за изчислението им от гледна точка на ресурси, проследимост и оптимално съотношение между сложност и обяснителна сила.[4]

K-най-близки съседи

Алгоритъмът k-най-близки съседи има дълга история и разнообразни приложения за класификация, което се дължи както на относителната му простота, така и на сравнително добрите резултати, до които води. Главното при него е, че това е класификационен алгоритъм, използващ вече познатите класове, които се намират в достатъчно близка околност до дадено наблюдение, за да определят принадлежността и на самото наблюдение. Основният параметър тук е броят най-близки наблюдения (съседи), които да се използват при решаването на тази задача.

По-конкретно, при нужда да се класифицира дадено наблюдение x_0, алгоритъмът търси k на брой точки за обучение $x_{(r)} = 1, 2, \ldots, k$ с възможно най-малко разстояние от x_0 и определя принадлежността на x_0 към даден клас спрямо най-често срещания клас сред наблюденията $x_{(r)}$. При реални стойности на обяснителните променливи за мярка за разстояние може да се използва евклидовото разстояние между наблюденията:

$$d_{(i)} = \left\| x_{(i)} - x_0 \right\|. \tag{4}$$

Това разстояние обикновено се изчислява, след като обяснителните променливи са стандартизирани (със средно аритметична от 0 и стандартно отклонение от 1) – класификацията се извършва чак след това [5]. Макар този алгоритъм да е компактен и сравнително интуитивен за разбиране и прилагане, той може да бъде успешен в редица ситуации, като резултатите му са особено добри в случаите с нерегулярни граници между различните класове на данните или когато всеки клас има редица ясно разграничени прототипи.

Бейсови класификатори

С помощта на основни идеи от бейсовата статистика може да се изведе и алтернативен класификатор, който също се използва често в научните изследвания и в практиката – т. нар. наивен бейсов класификатор. При него отново задачата за разпределение към определена група се решава чрез конструиране на вероятностни разпределения с помощта на бейсовата теорема. По-конкретно

[4] За повече детайли относно статистическите особености и характеристиките на невронните мрежи вж. Ripley & Hjort (1996).

[5] Относно изчислението и статистическите свойства на този алгоритъм вж. по-подробно Peterson (2009) и Hastie et al. (2005).

от интерес е условното разпределение на наблюденията y_i спрямо класовете C_k, имайки информация за поредицата от обяснителни (независими) променливи x_i. Допускайки, че наблюденията са независими едно от друго, съвместното условно разпределение може да бъде описано по следния начин:

$$(5) \qquad p(C_i|x_i) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

В този случай със Z е отбелязан израз за скалиране, като $Z = p(x_i)$. След като разпределенията са изведени от алгоритъма, класификационният проблем е завършен след дефиниране на правило за определяне на принадлежност. Едно от най-често използваните правила е да се приеме класът с най-висока вероятност – по този начин наблюдение y_i се определя за принадлежащо към клас C_k при следното условие:

$$(6) \qquad y_i = \operatorname{argmax}_k p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

Наивният бейсов класификатор изисква значително по-малко изчислителни ресурси, отколкото алтернативни алгоритми за машинно самообучение, а често има и съпоставима с тях прогностична точност. Макар основното му допускане, че обяснителните променливи са независими една от друга, да е рядко изпълнено на практика, това не води до значително влошаване на неговите резултати. Допълнително предимство на този подход е, че той може лесно да бъде изчислен и върху много големи масиви от данни, което в редица случаи го прави подходящ за приложение в практиката[6].

Дървета на решения и случайни гори

Дърветата на решенията представляват алтернативен модел за моделиране на задачи, свързани с разпознаване на различни класове. Те водят началото си от класическите теории за взимане на решения и впоследствие са припознати като полезен инструмент за класификация в сферата на машинното самообучение. Използвайки масив от тестови данни, дърветата избират най-добрия класификатор между набор от обяснителни променливи, като този процес тече итеративно. Първоначално при първия възел алгоритъмът избира променливата, която най-добре различава класовете един от друг, и оптималната ѝ стойност за класификация. След това задачата се разклонява спрямо стойността на тази променлива и на новите възли отново се избира оптималната променлива и нейната стойност, като това води до нови разклонения. При достигането на решение графичното представяне на резултатите наподобява изключително много на обърнато дърво, откъдето идва и неговото име. По същество този алгоритъм изследва условни вероятности за принадлежност към даден клас при наличие на определени характеристики (стойности на обяснителни променливи).

[6] За повече детайли извън това кратко описание вж. класическата работа на Lewis (1998).

Нека допуснем, че се работи със задача за класификация между *m* класа, за които има N_m наблюдения в регион R_m. В този случай вероятността на наблюдение y_i да принадлежи към клас C_k в точка *m* е:

(7) $$p_{mk} = \frac{1}{N_m}\sum_{x_i \in R_m} I(y_i = C_k).$$

Целта на алгоритъма е да раздели региона R_m с помощта на най-добрата обяснителна променлива x_i, така че да може да прогнозира класа C_k на наблюденията y_i. По този начин всяко наблюдение във възел *m* се класифицира като принадлежащо към преобладаващия клас в този възел:

(8) $$C(m) = argmax_k(p_{mk}).$$

Допълнителни класификационни възли (точки) се създават дотогава, докато не се достигне някакъв предварително зададен брой или друг поставен критерий за прекратяване на изчисленията. Така създадените дървета за решения могат да бъдат използвани в широк набор от приложни полета, където е необходимо да се разграничат различни типове наблюдения (класове), вкл. и в областта на управлението на операционните рискове[7].

За разрешаването на основните проблеми на дърветата на решенията като силна вариация и прекомерно нагаждане може да се създаде сборен модел от тях. Събирайки определен брой дървета на решенията, те може да се комбинират в общ ансамблов модел – т.нар. случайна гора. При обучението на този модел се избират първоначално случайни извадки от данни и характеристиките им и на тяхна база се създава набор от дървета. Алгоритъмът най-напред избира извадка от изследваните данни от *b = 1* до *B*, както и определен брой случайно избрани обяснителни променливи. На базата на тези данни и променливи се изчислява дърво на решенията, което прави оптимална класификация във възел *m* с независимите променливи, изтеглени спрямо извадката *B*.

Процесът по оценка на параметрите на дървото спира при достигане на желания брой крайни възли. Това се повтаря предварително зададен от потребителя брой пъти и тези дървета се комбинират в един ансамбъл $\{T_b\}_1^B$, откъдето идва и наименованието „гора“. Както при невронните мрежи, случайните гори също могат да бъдат използвани за моделиране на целеви променливи, които са както дискретни, така и продължителни. При последните правилото за взимане на решения за стойността на дадено наблюдение (f^B_k) е, както следва:

(9) $$f_k^B = \frac{1}{B}\sum_{b=1}^{B} T_b(y).$$

[7] За допълнителни детайли относно този алгоритъм и неговите статистически характеристики вж. основното изследване на Breiman et al. (1984), което ги дефинира в съвременния им смисъл.

В случаите на класификация тя се определя спрямо мнозинството на изборите, направени от дърветата, или:

(10) $$C_k^B(y_i) = majority.vote\{C_b(y_i)\}_1^B.$$

Моделите на случайна гора са сравнително по-лесни за интерпретация, отколкото алтернативни алгоритми като невронни мрежи или машини с подкрепящи вектори, което ги прави предпочитани в някои ситуации. В допълнение тяхната прогностична точност е изключително добра и в много случаи надминават своите конкуренти по обяснителна сила и по качество на класификацията (вж. Gerunov, 2019, където този резултат е потвърден и при експериментални данни за икономически избор). Като трето предимство трябва да се посочи и това, че те изискват сравнително малък брой параметри при обучението им, което го улеснява значително. Сред недостатъците се откроява фактът, че понякога тези модели изискват значителна изчислителна сила по време на етапа на обучение и затова невинаги са подходящи в случаите, в които са необходими постоянни обновявания върху големи масиви от реалновремеви данни. Все пак случайните гори имат сериозни предимства и се налагат като един от водещите методи за анализ[8].

Машини с подкрепящи вектори

Машините с подкрепящи вектори са класификационни модели, които водят началото си от сферата на машинното самообучение (Cortes & Vapnik, 1995). При дадени класове на наблюденията (например два класа – рисков и нерисков) те се стремят да намерят оптималната класификация, като изчисляват оптималната хиперравнина (*m*-мерна равнина) по средата на най-голямото разстояние между най-близките точки на различните класове. Векторите, съставени от граничните точки в това пространство, се наричат „подкрепящи вектори“, откъдето произлиза и наименованието на това семейство от алгоритми. По същество параметрите на класификационния алгоритъм се оценяват чрез решаване на задачи на квадратичното програмиране.

По-усъвършенстваните машини с подкрепящи вектори могат да проектират данни с краен брой измерения върху равнини с повече измерения и да осъществят класификацията в тези равнини. По-нататък накратко е представен случаят на използване на алгоритъма за линейна класификация.[9] Обясненителните променливи в дадена задача отново се обозначават с x_i, а зависимата (целева) променлива с *y*, като целта е да се илюстрира използването на линейна машина с подкрепящи вектори за решаването на задачата. Дефинираме хиперравнина по следния начин:

[8] За допълнителна информация и повече детайли за този подход вж. основополагащата статия на Breiman et al. (2001).

[9] За по-задълбочено представяне, статистически особености и други варианти на машини с подкрепящи вектори вж. Hastie et al. (2005).

(11) $$\vec{w} \cdot \vec{x} - b = 0.$$

Тук с $\vec{w}$ е обозначен нормалният вектор към хиперравнината, а *b* е параметър. Целта на оптимизационната задача е да се намерят най-големите разстояния между подкрепящите вектори, т.е. да бъде минимизирано $\|\vec{w}\|$ при дадените ограничения:

(12) $$\begin{aligned} &\min\|\vec{w}\|, \\ &s.t.:\ y_i(\vec{w} \cdot \vec{x} - b) \geq 1. \end{aligned}$$

Решението на тази задача е и оптималният класификатор в машината с подкрепящи вектори. Това представяне е само за най-опростения случай. По-развитите варианти на този алгоритъм позволяват класификация между множество класове, с което превръщат метода в популярен избор за моделиране на широк клас от ситуации, вариращи от потребителски избор до случаи на операционни рискове. Макар машините с подкрепящи вектори да са с отлична прогностична точност при широк клас от проблеми, те имат и някои ограничения. Преди всичко интерпретацията на резултатите им е силно затруднена, като често е трудно те да бъдат разбрани от неспециалисти.

Критерии за точност на прогнозата

Изборът на оптимален класификационен алгоритъм в дадена задача често е предизвикателна задача. От една страна, е важно алгоритъмът да има добра прогностична сила, като правилно класифицира значителна част от подадените наблюдения. От друга страна, грешките при класификацията често носят различна стойност – например кредитополучател, който неправилно е класифициран като неблагонадежден, е пропусната полза, докато такъв, който е неправилно класифициран като благонадежден, е реализирана загуба. В този смисъл е важно да се обърне внимание не просто на общата точност на класификацията, но и на по-детайлизирани индикатори за качествата на алгоритъма като специфичност, чувствителност, F-статистиката, капа-статистиката и др.

Мерките за точност на класификацията показват какъв процент от наблюденията са класифицирани коректно и при какъв процент е определен грешен клас. Общата мярка за точност на класификацията показва каква част от прогнозите на модела са верни и каква – грешни (Матеев, 2016). На практика често има значение и в каква посока са грешките, поради което се налага изчисляването и на по-детайли индикатори за коректност на класификацията.

В матрицата на класификацията (Кабакчиева, 2012; Семерджиева и др., 2013), представена в табл. 1, се описват две измерения – от една страна, реален наблюдаван клас, а от друга, класификация от дадения модел. При съвпадение между реален и прогнозиран клас 1 се работи с верни положителни наблюдение (True Positives, *TP*), докато при съвпадение при отрицателен клас 0 между прогноза и реалност става дума за верни отрицателни наблюдения (True Negatives,

TN). При грешно прогнозиран нулев клас се отчитат грешни отрицателни, а при грешно прогнозиран първи клас – грешни положителни наблюдения.

Таблица 1

Обща матрица на класификацията

		Реален клас	
		1	0
Прогнозен клас	1	Вярно прогнозиран клас 1, *TP*	Грешно прогнозиран клас 1, *FP*
	0	Грешно прогнозиран клас 0, *FN*	Вярно прогнозиран клас 0, *TN*
	Общо	Клас 1, *P*	Клас 0, *N*

На базата на тези съотношения може да се дефинира поредица от индикатори за прогностичната точност на даден класификационен модел (Fawcet, 2004): общата балансирана прогностична точност, прецизността, чувствителността, специфичността, общият процент грешно прогнозирани наблюдения. За оценяване на прогностичната точност на класификационния алгоритъм понякога се използва и F-мярката, която се дефинира по следния начин:

$$F = \frac{2}{\frac{1}{Pr}+\frac{1}{Ss}} = \frac{2}{\frac{TP+FP}{TP}+\frac{P}{TP}} = \frac{2TP}{(TP+FP+P)}. \quad (13)$$

Сред индикаторите за качество на даден прогностичен модел трябва да се спомене и капа-статистиката (Carletta, 1996), която измерва степента на съгласуваност между реалния и прогнозирания клас и получава стойности в интервала между 0 и 1. Стойност на капа от 1 означава пълна съгласуваност между прогнозата и наблюдавания клас, т.е. 100% прогностична точност, а стойност нула – 0% прогностична точност.

Като алтернативна мярка за качеството на даден класификатор може да се използва площта под т.нар. ROC крива (Walter, 2005), или крива на работната характеристика. Пространството на работната характеристика е двумерно пространство, което показва как даден класификатор се представя спрямо пропорцията на реалните положителни наблюдения (чувствителността) и пропорцията на грешните положителни наблюдения. Първото е мярка за ползата от дадения класификатор, а второто – за цената, която се налага. Точковата класификация в резултат от приложението на даден алгоритъм може да бъде използвана, за да се изчислят чувствителността и пропорцията на грешно класифицирани наблюдения (вж. табл. 1). Тези два индикатора задават координатите на точката на алгоритъма в ROC пространството.

Много от класификационните алгоритми извеждат вероятностно разпределение (шанс) за това дадено наблюдение да принадлежи към даден клас или изчисляват продължителна тестова статистика. Това позволява в ROC пространството да се покажа не просто точковото представяне на алгоритъма, но цяла крива, отчитаща резултатите от него при различни стойности на параметър или тестова статистика. Тази крива е именно кривата на работната харак-

теристика, а площта под нея е мярка за това доколко е успешна класификацията (Fawcet, 2004). Площта под кривата (*AUC*) отчита замяната между генерирани ползи от класификатори и допуснати грешки и дава единен индикатор за сравняване между алтернативни класификационни модели (Walter, 2005; Tanwani et al, 2009). Стойността на тази площ се изменя в интервала между 0 и 1. Трябва да се подчертае, че диагоналът в пространството има $AUC = 0.5$, поради което полезните класификатори се характеризират с $AUC > 0.5$.

Наличието на широк набор от различни индикатори за оценка на класификационните модели предполага известна свобода за анализаторите при окончателната селекция. От една страна, това позволява известна гъвкавост, но от друга, затруднява автоматизацията на процеса. За да бъде осъществена такава автоматизация, е необходимо или високо ниво на съгласие между алтернативните индикатори, което да даде еднозначен избор на оптималния модел, или една обобщаваща метрика, която може да бъде използвана като подходящ критерий за избор. Според редица автори (Fawcet, 2004; Walter, 2005; Tanwani et al, 2009) по този начин може да бъде използвана площта под кривата на работната характеристика.

Приложение при операции с кредитни карти

Рискът за издаващата кредитни карти организация при операции с тях може да се изследва, като се използват данните от проучване на Yeh и Lien (2009)[10], с които те сравняват прогностичната точност за вероятност от неплатежоспособност, прилагайки 6 алтернативни метода. Авторите откриват, че изкуствената невронна мрежа има най-добри прогностични резултати. Предвид ограничения брой тествани алгоритми и наличието на нови развития през последните десет години, тези данни могат да се използват за тестване на широк набор от алтернативни алгоритми. Масивът от данни съдържа 30 хил. наблюдения за клиенти, използващи кредитни карти в Тайван. Данните са разделени на няколко групи, като включват не само основни демографски характеристики на респондента, но и основни характеристики на кредитната карта, вкл. текущи баланси и натрупани задължения по нея.

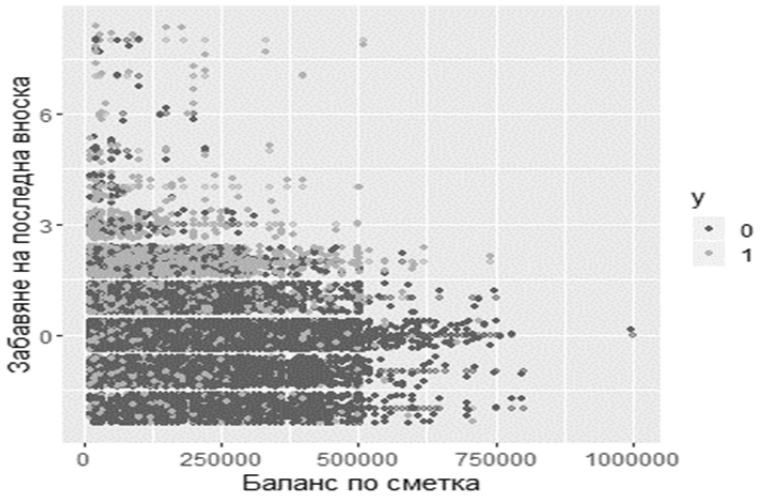
Основният бизнес проблем тук е да се избегне неплатежоспособност от страна на клиента. Поради това целевата променлива е именно статусът на кредитната карта, който показва дали дългът по нея е погасен (клас 0), или не е (клас 1). Необичайното поведение се проявява в необслужването на дълговете (клас 1) – то е както по-рядко, така и нежелано от гледна точка на издаващата институция. Това е и реализацията на операционния риск, който трябва да се управлява. Необходимо е да се отбележи, че в данните се съдържа информация

[10] Анализираните тук данни са предоставени от авторите им за свободно ползване с цел репликация на получените резултати и провеждане на вторични анализи върху други проблеми. Тези отворени данни са достъпни на редица места, като за нуждите на нашето изследване са свалени от Хранилището за данни за машинно самообучение на Университета в Калифорния Ървай (UCI Machine Learning Repository), https://archive.ics.uci.edu/ml/index.php.

за статуса на кредитната карта (обслужвана или не), но не се знае каква е причината за този статус. Особено голям риск са измамите с кредитната карта, когато неплащането на дълга по нея е резултат от действията на злонамерени ползватели, които съзнателно прибягват до тази възможност. Друг вероятен сценарий е при добронамерени клиенти, които просто нямат необходимата ликвидност да погасят дълга си. В този смисъл целевата променлива отчита едновременно операционния риск от измами и ликвидния риск на контрагента, като често е трудно на базата на данните да се определи кой от двата риска се е реализирал. От гледна точка на организацията това е второстепенно – независимо от характера на причината, тя реализира загуби в един и същи обем и мащаб и трябва да управлява рисковете по сходен начин на базата на разполагаемата информация.

Фигура 1

Връзка между месеци забавяне на последната вноска и баланса по сметката в данните за операции с кредитни карти



След първоначалното запознаване с данните трябва да се разгледат подетайлно променливите, които имат силна връзка както помежду си, така и с целевата (зависима) променлива. Фиг. 1 илюстрира зависимостта между баланса по сметката и броя месеци забавяне на последната вноска по натрупаните задължения. Очаквано, връзка между двете е отрицателна – клиентите с по-висок баланс по сметка натрупват по-малки забавяния. Съпоставяйки тези връзки с целевата променлива *y*, се очертава ясно изразена вероятност даден клиент да не обслужи дължимия баланс – клиенти с големи забавяния и малки баланси са най-вероятните носители на риска. В цитираното проучване Yeh и Lien установяват, че при баланс по сметка над 500 хил. нови тайвански долара

почти не се наблюдават случаи на необслужване. Същевременно при забавяне на последната вноска с повече от 2 месеца рязко се увеличава вероятността за прекратяване на плащанията по дължимата сума.

Характеристиките при класификация на моделите на логистична регресия и линеен дискриминантен анализ са представени в табл. 2.

Таблица 2

Класификационна матрица на разглежданите методи, използващи данни за операции с кредитни карти

	Логистична регр.	Линеен дискр. анализ	k-най-близки съседи	Наивен бейсов класификатор	Невронна мрежа	Случайна гора	Машина с подкрепящи вектори
Прогностична точност, %	81.1	81.1	76.8	79.3	77.9	81.9	81.0
Капа статистика,%	28.0	28.5	11.4	17.7	0.0	37.1	28.9
Долна граница на точността, 95%	80.6	80.1	75.7	78.2	76.8	80.9	80.0
Горна граница на точността, 95%	81.6	82.1	77.9	80.3	78.9	82.9	82.0
Безусловна вероятност, %	77.9	77.9	77.9	77.9	77.9	77.9	77.9
Значимост на разлика между класификация на модела и случайност	p < 0.0005	p < 0.0005	p = 0.977	p = 0.004	p = 0.507	p < 0.0005	p < 0.0005
Чувствителност, %	97.4	97.1	94.7	97.4	100.0	94.9	96.8
Специфичност, %	23.9	24.8	13.9	15.4	0.0	36.1	25.6
Правилно прогнозиран клас 1, %	81.8	82.0	79.5	80.2	77.9	83.9	82.1
Правилно прогнозиран клас 0, %	72.1	70.9	42.6	63.1	0.0	66.9	69.2
F-статистика	0.889	0.889	0.864	0.880	0.876	0.891	0.888
Площ под ROC-крива	0.610	0.615	0.544	0.564	0.500	0.655	0.612

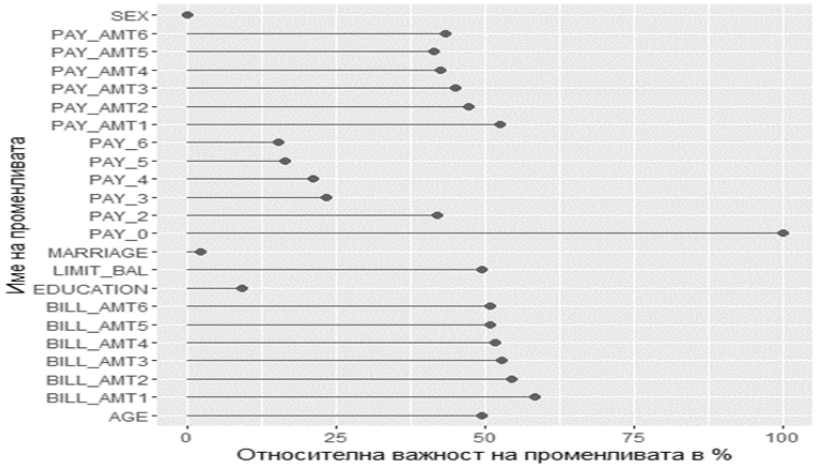
Прогностичната точност и на двата модела е сравнително висока, като достига над 81% както в обучаващите (непоказани), така и в тестовите извадки. И в четирите разгледани случая моделите генерират прогнози, които са статистически значимо по-добри, отколкото класификацията въз основа на безусловната вероятност (при нея прогностичната точност е 77.9%). Поради изключително близките стойности между представените алгоритми не може да се отчете единият или другият като по-добър в класификационната задача.

Макар общата прогностична точност на *случайната гора* да спада драстично спрямо обучаващата извадка (непоказана) поради свръхнагаждане към нея, при този алгоритъм точността остава най-висока. Много близък по успех алгоритъм е *машината с подкрепящи вектори* с обща прогностична точност от 81.0%. И при двата алгоритъма се наблюдава значително по-добра прогностична точност при клас 1 (в интервала 82-84%) и доста по-слаба при прогнозирането на нулевия клас (в интервала 67-69%). В общи линии това е адекватен баланс, тъй като основната цел на задачата е точното определяне на положителния рисков клас (неплащане) и управлението му. За да се случи това, се плаща цената на погрешните прогнози за нулеви класове, т.е. наблюдава се сравнително по-ниска прогностична точност при нулевия клас. При тестовата извадка както *невронната мрежа*, така и *kNN-алгоритъмът* не са значимо различни от класификацията, базирана на безусловна вероятност, което ги прави непродуктивни за целите на задачата.

За разлика от тях наивният бейсов класификатор е статистически значим, но с много по-слаба класификационна точност, отколкото двата най-добри алгоритъма. Както при обучаващата, така и при тестовата извадка най-голяма площ под кривата на работната характеристика има случайната гора, следвана от линейния дискриминантен анализ, т.е. тези два алгоритъма са най-подходящи за решаване на съответната задача. Фиг. 2 илюстрира относителната важност на факторите в най-добрия модел – случайната гора.

Фигура 2

Важност на различните обяснителни променливи в модел на случайна гора



Прави впечатление, че най-важната променлива е статусът на обслужването на дължимия баланс в настоящ период (PAY_0), следвана от двете групи променливи за размера на този дължим баланс (съответно PAY_AMT1 и PAY_AMT2), както и за размера на платените вноски по него и техните лагове за 5 месеца назад (BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6). Демографските променливи – брачен статус (MARRIAGE), образование (EDUCATION) и пол (SEX), се оказват с най-малка важност, по което случайната гора има значителна прилика с резултатите от линейния дискриминантен анализ.

Приложение при управление на кредитен риск

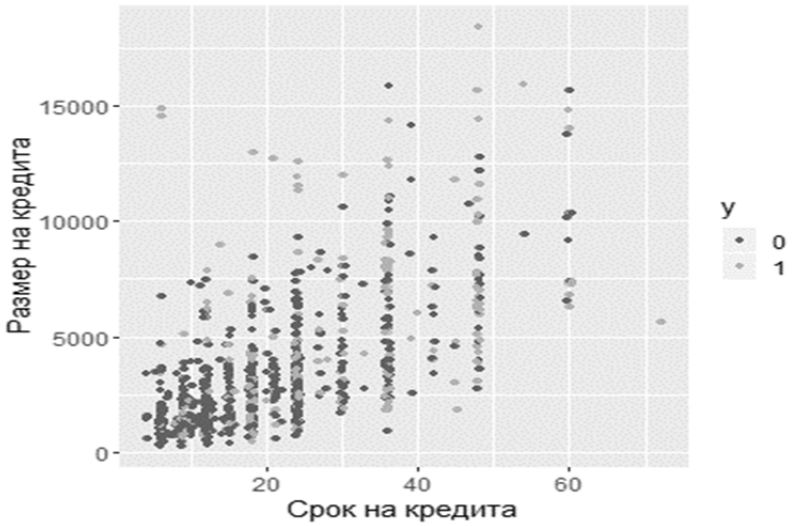
От бизнес гледна точка проблемът с управлението на риска при отпускане на кредити е по същество проблем на несъвършената информация. Даден заемодател или финансов посредник (например банка) преценява дали да отпусне кредит на определен заемополучател, като не знае предварително вероятността за обслужването му. Целта на управлението на риска е да се премине от решения, базирани на безусловни вероятности (обща вероятност

за обслужване), към персонализирани решения, основани на условни вероятности, т.е. дали разглежданото физическо или юридическо лице ще обслужи конкретния кредит. За целта се прилагат класификационни модели, обучени на базата на предишни данни, и новите наблюдения (потенциални клиенти) се класифицират като благонадеждни или не.

За да се анализира този проблем, могат да се използват данните, предоставени от Hoffmann (1994) и проучени от Eggermont et al. (2004), които описват 1000 клиенти на немска банка и състоянието на техните кредити. От тях 700 са обслужвани („добри“), а 300 не са („лоши“). Това е и изследваната целева променлива, като нежеланото аномално рисково събитие е съответният кредит да е необслужван. Целевата променлива е кодирана съответно 0 за нормални добри кредити и 1 – при аномални необслужвани такива. Трябва да се подчертае, че разпределението на кредитите в генералната съвкупност вероятно е по-различно, тъй като 20% необслужвани кредити се смятат за прекомерно висок дял от гледна точка на финансовото здраве на организацията. Извън целевата променлива масивът от данни съдържа още 20 характеристики на кредитополучателя, които могат да се използват като независими променливи при оценка на статистическите модели и на алгоритми от сферата на машинното самообучение. Фиг. 3 представя по-детайлно връзката между срока на кредита и неговия размер.

Фигура 3

Необслужване на конкретен кредит спрямо неговия размер и срок



Представените данни говорят за устойчива връзка между срока и размера на кредитите, като прави впечатление, че огромната част от кредитите са за срок

до 24 месеца (2 години). Интересен факт е, че се открояват две ясно изразени тенденции. От една страна, процентът на необслужените кредити нараства с увеличение на техния срок след преминаване на една критична граница от около 3 години. От друга страна, краткосрочните кредити в особено големи размери почти винаги са необслужвани.

Прогностичните качества на двата модела са представени чрез изчисляване на съответните класификационни матрици на базата на обучаващата и тестовата извадки (табл. 3).

Таблица 3

Класификационна матрица на разглежданите методи върху данни за директен маркетинг

	Логистична регр.	Линеен дискр. анализ	k-най-близки съседи	Наивен бейсов класификатор	Невронна мрежа	Случайна гора	Машина с подкрепящи вектори
Прогностична точност, %	74.5	75.5	71.0	72.0	73.0	76.5	75.5
Капа-статистика,%	34.6	37.2	15.7	9.1	31.8	29.6	36.5
Долна граница на точността, 95%	67.9	68.9	64.2	65.2	66.3	70.0	68.9
Горна граница на точността, 95%	80.4	81.3	77.2	78.1	79.0	82.2	81.3
Безусловна вероятност, %	70.0	70.0	70.0	70.0	70.0	70.0	70.0
Значимост на разлика между класификация на модела и случайност	p = 0.093	p = 0.051	p = 0.412	p = 0.297	p = 0.199	p = 0.025	p = 0.051
Чувствителност, %	87.1	87.9	92.9	100.0	85.0	98.6	88.6
Специфичност, %	45.0	46.7	20.0	6.7	45.0	25.0	45.0
Правилно прогнозиран клас 1, %	78.7	79.4	73.0	71.4	78.3	75.4	79.0
Правилно прогнозиран клас 0, %	60.0	62.2	54.5	100.0	56.3	88.2	62.8
F-статистика	0.787	0.794	0.818	0.833	0.815	0.854	0.835
Площ под ROC-крива	0.661	0.673	0.564	0.533	0.661	0.724	0.668

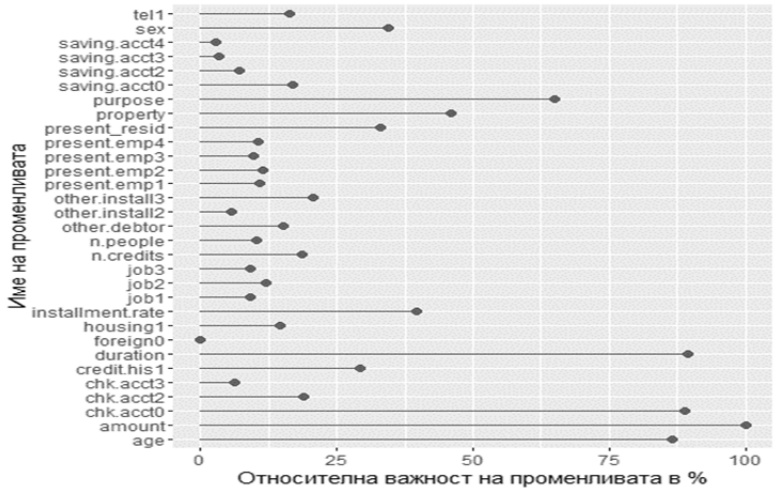
Разглеждайки представената класификационна матрица, могат да се направят различни изводи. Моделът на логистична регресия губи статистическа значимост при сравнението с класификация на базата на безусловна вероятност. Макар че класифицира правилно 74.5% от наблюденията, този модел може да постигне 70% точност, ако просто винаги поставя нулев клас. Това показва, че той не е полезен в решаването на задачата. Същевременно моделът на линеен дискриминантен анализ е по-добър от случайна класификация на базата на безусловната вероятност (p = 0.05) и класифицира правилно около 76% от наблюденията в тестовата извадка (79% в клас 1 и 62% в клас 0). Прогностичната точност не е идеална, но все пак е видна ползата му при решаването на класификационната задача при отпускане на банков кредит. Тази класификационна задача може да бъде решена и чрез прилагане на авангардни алгоритми от сферата на машинното самообучение. Отново случайната гора вероятно е достигнала до свръхнагаждане спрямо данните, което проличава от големия спад между прогностичната точност в обучаващата (непоказана) и тестовата извадка – от 94.3 до 76.5%. Въпреки това обаче случайната гора остава най-добрият прогностичен алгоритъм сред разгледаните и единственият, който е статистически значимо по-добър от безусловната класификация на нива от под

5% (p = 0.025). Машината с подкрепящи вектори клони към значимост на нива от 5% и показва втората най-добра прогностична точност сред разгледаните методи за машинно самообучение.

Останалите три алгоритъма не носят полезност при решаването на тази специфична задача при такъв масив от данни. Разглеждайки и резултатите за площ под кривата на работната характеристика, случайната гора се налага като най-добър класификационен алгоритъм, а линейният дискриминантен анализ – като втори най-добър. Проследявайки относителната важност на обяснителните променливи в случайната гора (вж. фиг. 4), се вижда, че най-важна променлива за модела е размерът на поискания кредит (amount), следвана от неговата продължителност (duration), разполагаемата сума в разплащателна сметка към настоящ момент (chk.acct0) и демографските характеристики на индивида – пол (sex) и възраст (age). Видима е и важността на целта на кредита (purpose), както и на притежанието на недвижима собственост (property).

Фигура 4

Важност на различните обяснителни променливи
в модел на случайна гора



Приложение при операции по директен маркетинг

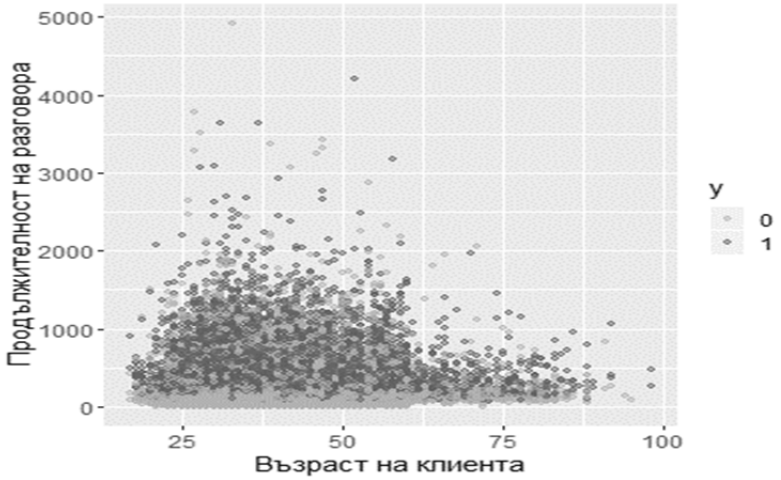
Moro et al. (2014) предлагат данни за провеждане на кампания за директен маркетинг от страна на голяма португалска банка. Базата се състои от данните, с които банката разполага за своите клиенти към момента на контакт с тях – 20 основни характеристики (демографски, финансови и отношения с банката). Извън тези 20 обяснителни (независими) променливи базата данни съдържа и индикатор за успех (*y*) на кампанията. Ако клиентът е избрал да приеме предложението за промоционален срочен депозит, зависимата променлива е кодирана като 1, а ако е отказал – като 0. Наблюденията в базата данни са 41 181,

но авторите тестват едва 4 класификационни алгоритъма върху тези данни, като най-добрият е невронна мрежа. Проблемът, който проучват, е по какъв начин банката може да прогнозира при кои клиенти вероятността да приемат предлагания депозит е висока, за да насочи маркетинговите си усилия към тези с най-голям потенциал. В такъв смисъл основната задача е да се класифицира всяко от наблюденията към съответен клас 0 или 1. Целевата променлива е индикаторът за успех (*y*), а нейната желана стойност е 1. Подобен тип задачи са особено подходящи за използване на класификационни алгоритми, като последователно се разглеждат описаните класически статистически алгоритми, както и по-модерните от сферата на машинното обучение.

Най-голяма положителна корелация между индикатора за успех и обяснителна променлива се наблюдава с продължителността на разговора, а корелацията има известен ефект и върху използвания канал на комуникация (вж. фиг. 5).

Фигура 5

Връзка между продължителност на разговора, възраст на клиента и успешна продажба



От фиг. 5 се вижда, че основните целеви сегменти на маркетинговата кампания са концентрирани върху лицата между 25- и 50-годишна възраст. Същевременно независимо от възрастта по-голямата вероятност за успешна продажба на даден клиент е свързана с продължителността на разговора.

Този казус може да бъде формално моделиран с помощта на класическите статистически методи, препоръчани в литературата. Както тук, така и по-нататък за оценката (обучението) на моделите се използва случайна обучаваща извадка от данни (80% от първоначалния масив), а за тестването на моделите и изчисляването на метрики за тяхната прогностична точност – тестова извадка (20% от първоначалния масив) (вж. табл.4).

Таблица 4

Класификационна матрица на разглежданите методи върху данни за директен маркетинг

	Логистична регр.	Линеен дискр. анализ	k-най-близки съседи	Наивен бейсов класификатор	Невронна мрежа	Случайна гора	Машина с подкрепящи вектори
Прогностична точност, %	91.2	91.3	90.8	88.7	90.1	91.5	90.4
Капа-статистика,%	47.6	53.3	49.9	0.0	27.8	55.3	38.4
Долна граница на точността, 95%	90.6	90.7	90.2	88.0	89.4	90.9	89.8
Горна граница на точността, 95%	91.8	91.9	91.4	89.4	90.7	92.1	91.1
Безусловна вероятност, %	88.7	88.7	88.7	88.7	88.7	88.7	88.7
Значимост на разлика между класификация на модела и случайност	p < 0.0005	p < 0.0005	p < 0.0005	p = 0.509	p < 0.0005	p < 0.0005	p < 0.0005
Чувствителност, %	97.3	96.2	96.1	100.0	99.0	96.0	97.9
Специфичност, %	42.8	53.3	49.8	0.0	20.2	56.5	31.9
Правилно прогнозиран клас 1, %	93.1	94.2	93.8	88.7	90.7	94.6	91.9
Правилно прогнозиран клас 0, %	67.1	63.8	61.6	0	71.1	64.1	65.6
F-статистика	0.951	0.951	0.949	0.940	0.947	0.953	0.948
Площ под ROC-крива	0.700	0.734	0.730	0.500	0.726	0.748	0.636

От представеното в табл. 4 прави впечатление, че и двата традиционни подхода се характеризират с добра обща прогностична точност, като не се отчитат сериозни разлики между обучаващата и тестовата извадка. Общата прогностична точност е 91%, а по-големи слабости се наблюдават при прогнозата на отрицателния клас 0 – при логистичната регресия точността е около 64%, а при линейния дискриминантен анализ – в порядъка на 62%. При всички случаи и двата оценени модели са подходящи за дадената задача, като тяхната класификация е статистически значимо по-добра от тази, дължаща се на шанса с точно ниво на значимост значително под 1%.

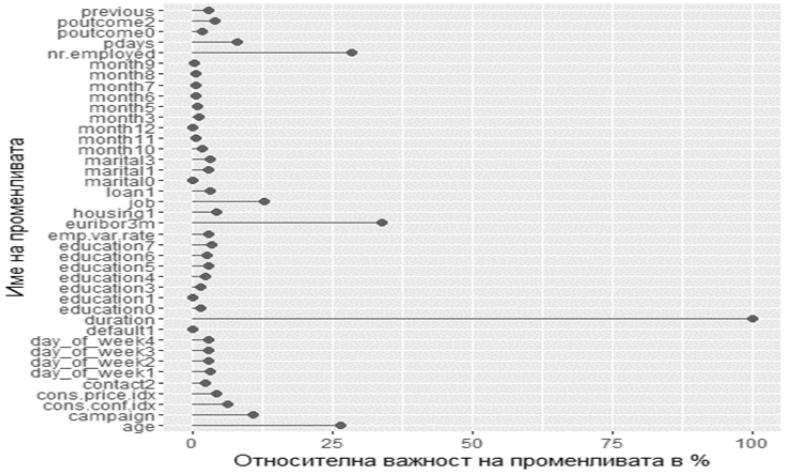
Логистичната регресия дава малко по-добри резултати предвид прогностичната точност в класа на неуспешните продажби (клас 0). Алгоритмите за машинно самообучение също имат доста добро представяне, като с най-висока обща прогностична точност е моделът на случайна гора. При сравнението между обучаващата и тестовата извадка се отчита леко влошаване в прогностичната сила на всички алгоритми, което е най-забележимо в случая на случайната гора. Тук спадът на точността е от 100 на 91.5%, но все пак това остава най-добрият от разгледаните алгоритми. Подобен спад показва нагаждане спрямо данните и насочва към потенциални рискове при захранване с други извадки от генералната съвкупност. Същевременно невронната мрежа, kNN-алгоритъмът и машината с подкрепящи вектори задържат приблизително същите стойности на общата прогностична точност. Невронната мрежа има и най-висока прогностична точност при отрицателния клас, докато при положителния всички алгоритми имат съизмерими стойности. Наивният бейсов класификатор продължава да показва най-слаби резултати – неговите прогнози на практика не са различими от шанса.

Представянето на основните типове класификационни алгоритми е измерено чрез площта под ROC-кривата. Отново се виждат отличните резултати при случайната гора (AUROC = 0.748) и при линейния дискриминантен анализ (AUROC = 0.734). Доброто представяне на случайната гора е видимо и в трите разгледани казуса, като този резултат не е необичаен (вж. Fernandez-Delgado et al., 2014; Gerunov, 2019).

Макар случайната гора да е с общ по-висок резултат, все пак има редица случаи, при които може да бъде избран и дискриминантният анализ. Това важи най-вече при анализ на особено големи масиви от данни при ограничени ресурси – за разлика от случайната гора линейният дискриминантен анализ е с много висока изчислителна ефективност и са му нужни значително по-малко ресурси. Останалите разгледани подходи имат по-слаби резултати, което показва, че тяхната полза за подобен тип задачи е сравнително по-малка. Невронните мрежи и машините с подкрепящи вектори демонстрират най-слабо представяне в тази конкретна задача със съответни стойности на площта под ROC-кривата от 0.636 и 0.5 при тестовата извадка.Най-важните за класификацията променливи са представени на фиг. 6.

Фигура 6

Важност на различните обяснителни променливи
в модел случайна гора



Продължителността на разговора (duration) е определяща за успешното приключване на сделката, последвана от индикаторите на обективната икономическа среда – брой заети (nr.employed) и тримесечен Euribor като индикатор на лихвените проценти (euribor3m). Демографските характеристики на потенциалния клиент като ниво на образование (променливи education0 до education7) и заетост (job) с изключение на възрастта (age) не са особено важни фактори, определящи вероятността за сключване на сделка.

*

Резултатите на изследваните пет алгоритъма от сферата на машинното самообучение – невронна мрежа, бейсов класификатор, kNN-алгоритъм, машина с подкрепящи вектори, случайна гора, са сравнени с представянето на логистична регресия и на линейния дискриминантен анализ. Те са приложени към три основни задачи – за определяне на вероятност за непогасяване на дълг по кредитна карта, за управление на кредитния риск в банковия сектор и за извеждане на вероятност за закупуване на промоционален финансов продукт. На базата на площта под кривата на работната характеристика се установява, че във всички случаи традиционните подходи на логистична регресия и на линеен дискриминантен анализ имат по-слаба прогностична точност спрямо най-добрия подход от сферата на машинното самообучение. Това очертава възможностите за прилагане на някои от представените тук методи при решаване на икономически задачи, които могат да допълнят, а в някои случаи и да заместят обичайните иконометрични инструменти.

Използването на нови методи предполага и нов подход към интерпретацията на изчислените модели. За разлика от стандартния подход с разглеждане на оценените регресионни коефициенти – техния размер, знак и статистическа значимост, при новите алгоритми това не е приложимо. При тях интерпретацията може се получава, като се оцени относителната важност на независимите променливи за достигане до точна прогноза. В този смисъл най-важните фактори могат да се тълкуват като основни двигатели на процеса, а тяхната относителна важност – като количествен индикатор за размера на ефекта.

Извън методологичните съображения разработката дава и конкретни насоки за подходящите инструменти при моделиране на икономически задачи. Оптималната прогностична точност се отчита при случайната гора и в този смисъл е целесъобразно алгоритъмът да се използва като основен или алтернативен подход за анализ на дискретен избор в стопански контекст. Като втори най-добър алгоритъм се откроява линейният дискриминантен анализ, който често има почти същата прогностична точност като случайната гора. Същевременно линейният дискриминантен анализ е много по-познат и изчислително оптимизиран алгоритъм, така че неговата оценка изисква значително по-малко изчислителни ресурси. Затова е подходящо в ситуации на ресурсни ограничения да се използва линеен дискриминантен анализ, тъй като е вероятно при него да се плати много ниска цена от гледна точка на загубата на прогностична точност. На другия край на спектъра са алгоритмите k-най-близки съседи и наивният бейсов класификатор – те систематично произвеждат слаби резултати. На базата на разгледаните извадки тези два алгоритъма не могат да се отчетат като особено полезни за моделиране на дискретен избор в стопански контекст.

Накрая ще завършим с въпроса, зададен от Fernandez-Delgado et al. (2014), дали в действителност има нужда от стотици различни алгоритми, за да бъдат решавани реалистични класификационни задачи. Представените тук

резултати показват, че различните подходи се характеризират със значителни разлики в прогностичната точност. Същевременно дори малки подобрения в точността на критични процеси като управлението на кредитния риск във финансовия сектор имат потенциала да доведат до значителни резултати предвид мащаба на тяхното приложение. В този смисъл използването само на един подход е недостатъчно, а търсенето и успешното намиране на оптимални модели сред множество потенциални алтернативи е ключово не само за икономическата теория, но и за стопанската практика.

Използвана литература:

Кабакчиева, Д. (2012). *Изследване на Data Mining модели за класификация* (дисертация за присъждане на ОНС „доктор"). С.: Институт по информационни и комуникационни технологии, БАН.

Матеев, С. (2016). *Оценка на методи за диагностика и прогнозиране – аналитични процедури и интерпретация на данните*. Нов български университет.

Семерджиева, В., Б. Георгиев, Ч. Дамянов (2013). Анализ на данни от диагностични тестове. *Научни трудове на УХТ*, *60*, с. 292-297.

Akinci, S., E. Kaynak, E. Atilgan, & Ş. Aksoy (2007). Where does the logistic regression analysis stand in marketing literature? A comparison of the market positioning of prominent marketing journals. *European Journal of Marketing*, *41(5/6)*, pp. 537-567.

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

Breiman, L., J. H. Friedman, R. A. Olshen & C. J. Stone (1984). Classification and regression trees. Belmont, CA: Wadsworth. *International Group*, *432*, pp. 151-166.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, *22(2)*, pp. 249-254.

Chandola, V., A. Banerjee & V. Kumar (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41(3)*, pp. 1-58.

Cortes, C. & V. Vapnik (1995). Support-vector network. *Machine Learning*, 20, pp. 1-25.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20(2)*, pp. 215-232.

Eggermont, J., J. N. Kok & W. A. Kosters (2004). Genetic programming for data classification: Partitioning the search space. *Proceedings of the 2004 ACM symposium on Applied computing*. ACM, March, pp. 1001-1005.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, *31*(1), pp. 1-38.

Fernández-Delgado, M., E. Cernadas, S. Barro & D. Amorim (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15(1)*, pp. 3133-3181.

Gerunov, A. (2019). Modeling Economic Choice under Radical Uncertainty: Machine Learning Approaches. *International Journal of Business Intelligence and Data Mining*, *14 (1-2)*, pp. 238-252.

Hastie, T., R. Tibshirani & J. Friedman (2013). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media. World classification problems? *The Journal of Machine Learning Research*, *15(1)*, pp. 3133-3181.

Hastie, T., R. Tibshirani & J. Friedman & J. Franklin (2005). *The elements of statistical learning: data mining, inference and prediction*. Springer Science & Business Media.

Hensher, D. A. & L. W. Johnson (2018). *Applied discrete-choice modelling*. Routledge.

Hofmann, H. (1994). *German Credit Data (Statlog)*. Institute for Statistic and Econometrics. University of Hamburg.

Hyman, M. R. & Z. Yang (2001). International marketing serials: a retrospective. *International Marketing Review*, *18(6)*, pp. 667-718.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. *In: European conference on machine learning*. Springer, Berlin, Heidelberg, pp. 4-15.

McFadden, D. (1981). Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*. US: Berkeley, pp. 198-272.

Moro, S., P. Cortez & P. Rita (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, pp. 22-31.

Omar, S., A. Ngadi & H. H. Jebur (2013). Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, *79(2)*.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4(2),* 1883.

Phua, C., V. Lee, K. Smith, & R. Gayler (2010). A comprehensive survey of data mining-based fraud detection research. *ArXiv preprint* arXiv:1009.6119.

Qiu, J., Q. Wu, G. Ding, Y. Xu & S. Feng (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, *2016(1)*, 67.

Ripley, B. D. & N. L. Hjort (1996). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.

Rousseeuw, P. J. & M. Hubert (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8(2)*, pp. 1-14.

Tanwani, A. K., J. Afridi, M. Z. Shafiq & M. Farooq (2009). Guidelines to select machine learning scheme for classification of biomedical datasets. *In: European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer, Berlin, Heidelberg, pp. 128-139.

Walter, S. D. (2005). The partial area under the summary ROC curve. *Statistics in medicine*, *24*(13), pp. 2025-2040.

Yeh, I. C. & C. H. Lien (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36(2),* pp. 2473-2480.

22.11.2019 г.