Ирена Стефанова*

СТАТИСТИЧЕСКО ОЦЕНЯВАНЕ НА РИСКА ПРИ БАНКОВО КРЕДИТИРАНЕ НА ГРАЖДАНИ

Представено е изграждането на логистичен регресионен скоринг-модел като статистически инструмент за определяне на вероятността даден кредитоискател да попадне в състояние на неизпълнение. Направено е изследване с реални данни за България. Описани са етапите на моделиране на регресионното уравнение: формиране на извадка от данни за моделиране; статистически анализ на данните по отношение на качество и пълнота; избор на подходящо логистично регресионно уравнение; анализ и оценяване на представянето на избрания регресионен модел. Използваните данни включват информация за граждани-кредитоискатели от банковия сектор у нас и са обработени с помощта на статистическия софтуер IBM SPSS Statistics v19.[1]

JEL: C25; C52; G21

Оценяването на кредитния риск е основна задача на управлението на бизнеса с банково кредитиране по отношение на разпределението и диверсификацията на риска, ценообразуването и връзката с доходността на кредитния продукт, измерването и оценяването на кредитните експозиции и техния дял в общия кредитен портфейл. Необходимостта от оценяване на кредитния риск нараства все повече през последните години, от една страна, основно поради бързото разпространение на различните кредитни продукти, а от друга – поради глобализацията в достъпността на тези продукти за все по-голяма част от участниците на пазара.

Развитието на пазарите и стремежът към постигане на все по-висока печалба тласкат инвеститорите към вземане на решения, които да им позволят поемането на по-голям кредитен риск, докато управляват своите експозиции в динамична, бързо променяща се среда.

Управлението на кредитния риск има два основни аспекта: първият е свързан с определяне на вероятността от попадане в състояние на неизпълнение на плащанията по кредита, а вторият – с установяване на частта, която би могла да се възстанови при условие, че е настъпило състояние на неизпълнение.

* Докторант в ИИИ при БАН, секция „Макроикономика”, renineykova@abv.bg

[1] Irena Stefanova. STATISTICAL RISK ASSESSMENT IN BANK LENDING TO CITIZENS.*Summary:* In this article, construction of a logistic regression scoring model is presented as a statistical tool to determine the likelihood a borrower to fall into a state of failure. The study is done with Bulgarian real data. The main stages of modeling regression equation are described: sampling data modeling; statistical analysis of the data in terms of quality and completeness; selection of an appropriate logistic regression equation; analysis and evaluation of the performance of the selected regression model. The used data include information about citizens who are borrowers in the banking sector in Bulgaria. The data were processed by means of the statistical software IBM SPSS Statistics v19.

През последните години, особено след финансовата криза от 2008 г., доброто познаване на кредитния риск и неговото управление се превръщат в основна необходимост и задължително изискване за банките. Ето защо банковият сектор в България изпитва силна нужда от развитие и използване на инструментариум за оценяване и анализ на вероятността от попадане на кредитоискателите в състояние на неизпълнение. Полезно е оценяването на тази вероятност да се извършва както в момента на кандидатстване за кредит, така и по време на неговия живот. Докато една част от банките у нас използват собствени модели за оценяване на кредитния риск при кредитиране на граждани, други взаимстват готови модели, разработени от банката-майка. И в двата случая, за да имат добро представяне, прилаганите модели за оценяване на кредитния риск при кредитиране на физически лица трябва да съответстват на икономическите и финансовите специфики в поведението на гражданите в страната.

Формиране на извадка от данни за моделиране

Първата стъпка при създаване на скоринг-модел е да се определи наборът от данни (извадката), на чиято основа се изгражда логистично регресионно уравнение. Изборът зависи преди всичко от типа на кредитния продукт, за който трябва да се построи съответният модел. Богатата гама от кредитни продукти, предлагани от банковия сектор в България, обуславя възможността за създаването на различни скоринг-модели за всеки от тях. Предпоставка за избор на различен набор от данни са специфичните особености в параметрите на кредитните продукти - например формиране на извадка само от кредити за текущо потребление на граждани или от жилищни и ипотечни кредити на граждани. Тъй като кредитоискателите се отличават един от друг според желания тип на кредитния продукт, за предпочитане е да се изграждат скоринг-модели за различните кредитни продукти, а не да се създава общ модел, обхващащ всички кредитни продукти на граждани.

Второ, необходимо е също да се специфицира видът на информацията, която трябва да се съдържа в извадката. Най-общо тя може да се обобщи в следните групи: социо-демографски данни за клиента (пол, образование, семеен статус, упражнявана професия и др.); информация от Централния кредитен регистър (ЦКР) за наличието и размера на общата кредитна експозиция на кредитоискателя; информация за наличие и размер на салдо по депозитни и/или спестовни сметки на клиента.

Трето, трябва да се определи времевият обхват на набора от данни. Този етап от формирането на извадката е един от най-важните и оказва силно влияние върху крайния резултат при изграждане на скоринг-модела за оценяване на вероятността от попадане в състояние на неизпълнение. За да се създаде регресионен модел с добро представяне, е необходимо данните, върху които той се базира, да бъдат възможно най-близко по време с тези, върху които ще се прилага полученият регресионен модел. Ако данните, върху които е

построен регресионният модел, са твърде отдалечени назад във времето, тогава има реална опасност създаденият регресионен модел да прояви слаба познаваемост и недобро представяне при определяне на вероятността от попадане в състояние на неизпълнение. Това може да се дължи на настъпили съществени промени в икономическата и финансовата среда, които не са обхванати като важни фактори в данните, върху които е построена логистичната регресия.

При скоринг-модели, изградени чрез логистична регресия, зависимата променлива приема най-често две алтернативни стойности: 1 - обозначава неизпълнение на задължението по кредита, и 0 – изпълнение на задълженията. Определянето на тези стойности заема основно място при дефинирането на времевия обхват на набора от данни за изграждане на скоринг-модела. Логистичната регресия моделира вероятността от попадане в състояние на неизпълнение за конкретен кредитоискател. За да се отчете дали даден кредитоискател е попаднал в състояние на неизпълнение, или не, трябва да се определи хоризонт от време, през което да се наблюдава появата на това състояние. Ето защо при определяне на времевия обхват на извадката от данни за моделиране е необходимо да се изберат данни назад във времето по такъв начин, че да е възможно конструирането на зависимата променлива. Времевият хоризонт за наблюдение на появата на състояние на неизпълнение зависи от конкретното събитие, чието предвиждане трябва да се моделира чрез логистичната регресия - понякога той може да е 3 или 6 месеца, но по-често се използва 12-месечен период.

Тук е формирана извадка от данни за физически лица, ползващи кредити за текущо потребление от банковата система в България. Извадката съдържа общо 72 920 разрешени кредита за период от януари 2013 г. до декември 2013 г. вкл. От една страна, времевият обхват на извадката от данни за моделиране е избран така, че да бъде осигурен 12-месечен период след датата на разрешаване, през който да се проследи поява на събитието „попадане в състояние на неизпълнение“ за всеки от разрешените кредити. От друга страна, времевият хоризонт на данните обхваща разрешени кредити за период от една година, а не няколко месеца. Това е направено с цел да се избегне появата на сезонност в данните, която може да настъпи, ако те обхващат 1 или 2 тримесечия. Например в дните около празници (новогодишни, великденски и т.н.) много често банките провеждат кампании по предлагане на кредитни продукти с по-приемливи параметри от тези, които обикновено предоставят. Нормално е по време на такива кампании потреблението на кредити да нараства спрямо обичайното темпо за конкретните кредитни продукти.

За всеки разрешен кредит за текущо потребление от извадката е проследено неговото състояние по отношение на редовното погасяване на месечното задължение в продължение на 12 месеца след датата на разрешаване. По този начин е създадена зависимата променлива Y, която приема стойност 1, ако плащането на дължимата месечна вноска по кредита е просрочено над 90

дни през изследвания 12 месечен период, и стойност 0 в противоположния случай. Според стойността на променливата Y случаите от извадката могат да се разделят в две групи: „недобри“ кредитополучатели –тези, за които Y приема стойност 1, и „добри“, за които стойността ѝ е 0.

При формирането на извадката много често се установява, че за някои от клиентите не може да се определи дали са „добри” или „недобри”, защото те не съществуват през целия изследван 12-месечен времеви хоризонт или тяхната кредитна история не е ясна. Такива случаи са отстранени от извадката. Формираната по този начин извадка от данни за кредитополучатели за периода януари 2013 – декември 2013 г. съдържа 69 473 (95,27%) „добри“ и 3447 (4,73%) „недобри“ клиенти.

Важен при формирането на извадката е подборът на независими променливи. В изграждането на логистична регресия могат да участват както независими променливи от категориен тип, така и такива от непрекъснат тип. Съществено при изграждане на логистично регресионно уравнение е това, че Y е категорийна променлива, която в случая е от бинарен тип (приема само две стойности - 0 или 1).

Тук формираната извадка за физически лица включва информация за кредитополучателите по следните независими променливи, разпределени в няколко групи от данни:

- социо-демографски данни за клиента - възраст; пол; образование; семейно положение; упражнявана професия; общ трудов стаж в месеци; трудов стаж в години при сегашния работодател; собствено недвижимо имущество; размер и източник на нетния месечен доход; брой лица в домакинството; брой лица в домакинството, получаващи доход; брой членове в семейството; брой притежавани моторни превозни средства; брой притежавани мобилни телефони; период (в години), през който лицето живее на посочения от него адрес;
- информация от Централния кредитен регистър за граждани - размер на общата кредитна експозиция на клиента; сведения за наличие на просрочена част, както и на просрочена над 90 дни част от общата му експозиция;
- исторически данни за поведението на клиента по отношение на взаимоотношението му с банковата институция, в която е кредитополучател - обща сума на салдото по депозитни сметки; обща сума на салдото по спестовни сметки; информация дали лицето е съдлъжник/поръчител по съществуващ кредит.

Представителността на данните, формиращи извадката, с която е извършен анализът, се гарантира от начина, по който са подбрани отделните случаи – това са всички възможни кредити, които съществуват поне 12 месеца след тяхното разрешаване. Това позволява направените изводи от проведените статистически тестове да се приемат за валидни и по отношение цялата популация. Въз основа на взаимовръзките между зависимата и независимите променливи, които ще бъдат открити с помощта на логистичната регресия, ще се моделира вероятността даден кредитополучател да попадне в състояние на неизпълнение

повече от 90 дни на своето кредитно задължение в рамките на 12-те месеца, следващи месеца, през който се извършва оценяването.

Статистически анализ на данните по отношение на качество и пълнота

Веднъж формирана, извадката от данни за физически лица, ползващи кредити за текущо потребление, подлежи на анализ за качество на данните. Получените резултати зависят до голяма степен от пълнотата и качеството на изходните данни - резултатите ще бъде толкова по-надеждни, колкото по коректни и налични са те. Ако първичните данни имат недобро качество, т.е. съществуват твърде много липсващи стойности или такива, чието съдържание е неясно, тогава постигнатият резултат би бил незадоволителен спрямо поставените в анализа цели.

Преди да се пристъпи към анализиране на липсващите стойности, трябва да се провери дали на избраните независими променливи е присвоена коректната измерителна скала. Например независимата променлива „възраст“ приема числови стойности от изброимо множество и поради това е коректно да се запише, че тя е метрирана непрекъсната променлива. Освен такива могат да се срещнат и метрирани променливи, които имат прекъснат (дискретен) тип (например брой лица в домакинството; брой издържани лица и т.н.).

В процеса на статистическото изследване на данните понякога се налага непрекъснатите признаци да се преобразуват в дискретни, т.е. да се трансформират в прекъснати чрез закръгляване или чрез промяна в мащаба на измерване. Например възрастта може да се представи в навършени години, т.е. като прекъсната метрирана променлива, която може да заема само целочислени стойности.

Независимите променливи, които нямат числови значения, са от неметриран (категориен) тип (например пол, семейно положение, образование, упражнявана професия и др.). Некоректно е по време на анализа на променливи от такъв тип да се присвои метриран тип, защото това би довело до некоректно интерпретиране на данните, а оттук - и до несъдържателни оценки в анализа. Ординалната ска́ла представлява компромисен опит за числово характеризиране на номинални признаци. В случаите на кодиране на променливи, чиито значения са измерени в ординална ска́ла, числовите стойности обикновено се подбират така, че да имат смисъл и като измерители на посоката на различие - например степента на завършено образование (начало, средно, средно специално, полувисше/колеж[2] и висше), чиито значения могат да се ранжират съответно с „0“, „1“, „2“, „3“ и „4“. Числовите кодове в този случай нямат практически смисъл като

[2] Степента „полувисше образование“ идва от предходни години и неин аналог днес е степента „колеж“. Тези две форми често присъстват заедно в описанието на степента на завършено висше образование в декларациите, които лицата попълват при кандидатстване за кредит, тъй като част от тях са придобили своето образование в предходни години.

количествени съотношения на различията между стойностите на променливата „образование“, т.е. числото 4 е два пъти по-голямо от числото 2, но това не означава, че висшето образование, което е кодирано с „4“, е два пъти по-голямо от средно специалното, кодирано с „2“.

Ако подбраните променливи се окажат недостатъчни за изследваното събитие, могат да се създадат нови чрез използване на някои математически операции (събиране, изваждане, умножение, деление), например съотношение на нетния месечен доход на клиента и някои други променливи от социо-демографски характер като:

●съотношение на дохода на клиента към броя на лицата в домакинството;

●съотношение на дохода на клиента към броя на лицата в домакинството, получаващи доход;

●съотношение на дохода на клиента към брой издържани лица в домакинството.

Статистическият анализ на данните по отношение на тяхното качество и пълнота включва също и изследване на описателните характеристики на всяка от избраните независими променливи (табл. 1). Сведенията за максимална, минимална, средноаритметична стойност, мерки за разсейване, както и за симетричност и върхова изтъченост на тези променливи дава първоначална представа за качеството на наличната информация в извадката от данни.

Таблица 1

Описателни характеристики на независимите променливи

Променлива	Мин. ст-ст	Макс. ст-ст	Средно-аритметична ст-ст	Станд.грешка на средно-аритметичната ст-ст	Стандартно отклонение
Възраст	18	80	53.145	0.06	16.133
Образование	--	--	--	--	--
Общ трудов стаж в месеци	0	720	286.291	0.563	152.012
Трудов стаж в месеци при последния работадател	0	57	7.801	0.026	6.985
Нетен месечен доход	0	5923	351.065	1.955	528.046
Променлива	**Дисперсия**	**Асиметрия**	**Ескцес**	**Медиана**	**Мода**
Възраст	260.265	-0.514	-0.85	58	64
Образование	--	--	--	--	1
Общ трудов стаж в месеци	23107.523	-0.308	-1.002	312	360
Трудов стаж в години при последния работадател	48.793	1.18	1.619	6	1
Нетен месечен доход	278832.299	40.433	3055.377	269	490

Информацията в табл. 1 показва, че липсват некоректни стойности в данните. Всички стойности са в приемлив диапазон според логическия и икономическия характер на изследваните величини. Според стойностите за асиметрия (Skewness) и ексцес (Kurtosis) може да се предположи, че независимите променливи не са нормално разпределени. За потвърждаване на това предположение може да се използва тестът на Колмогоров-Смирнов за проверка на нормално

разпределение (вж. табл. 2), който показва, че изследваните независими променливи имат разпределение, различно от нормалното. Значимостта на проведения тест на Колмогоров-Смирнов е със стойност, по-малка от 0,05, което означава, че нулевата хипотеза (няма разлика между разпределението на данните и нормалното разпределение) се отхвърля в полза на алтернативната (има разлика между разпределението на данните и нормалното разпределение).

Таблица 2

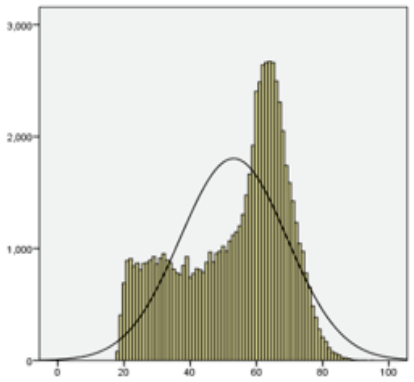
Тест на Колмогоров-Смирнов за нормалност на разпределението

Независима променлива	Колмогоров-Смирнов		
	Статистика (Statistic)	Степени на свобода (df)	Значимост (Sig.)
Възраст	.132	72743	0.000
Трудов стаж в месеци при последния работодател	.132	72743	0.000
Общ трудов стаж в месеци	.115	72743	0.000
Брой членове в семейството	.349	72743	0.000

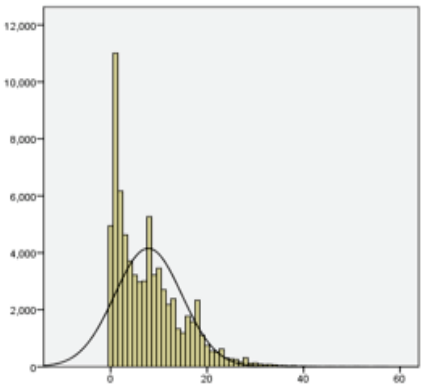
Липсата на нормално разпределение в изследваните независими променливи може да се установи и чрез тяхното честотно разпределение (хистограма). На фиг.1 и 2 е представено сравнение между честотното разпределение на две от независимите променливи - „възраст“ и „“трудов стаж в години при сегашния работодател“, и теоретичната крива на нормалното разпределение.

Честотно разпределение

„Възраст“ **Фигура 1**



„Трудов стаж в години при настоящия работодател“ **Фигура 2**



Липсата на нормално разпределение на данните в избраната извадка не позволява използването на линеен дискриминантен анализ като статистическа

техника за изследване на наличие на взаимовръзки между тези променливи. В случаи, в които не може да се приложи такъв анализ, е подходящо да се избере прилагането на логистична регресия.

Почти винаги при работа с бази от данни се наблюдава наличие на липсващи стойности. Техният анализ има съществено значение при построяването на модели на взаимовръзки между изследваните величини. Извадка с висок процент на липсващи стойности крие голяма опасност от постигане на нереални резултати в моделирането на икономически събития. Във връзка с това, от една страна, може да се предприеме отстраняване на случаите, за които е установено, че имат липсващи стойности по някои от изследваните променливи. Такива ситуации крият риск от силно намаляване на размера на формираната първоначално извадка от данни, ако делът на липсващите стойности е твърде голям, а това поражда проблем с представителността на данните. От друга страна, при наличие на липсващи стойности може да се предприеме заместването им с подходяща според данните стойност. Рискът в тази ситуация се изразява в това, че ако съществена част от данните се замести с една и съща стойност по даден показател, тогава изкуствено се създава нереално съществуване на данни според този показател. Ето защо при наличие на липсващи стойности в изследваните данни възниква въпросът кой от двата варианта да се избере. В отговор на този въпрос в практиката са се наложили различни подходи. При част от тях се отстраняват променливи величини с липсващи стойности над 3%. Този подход е рестриктивен и е подходящ, когато във формираната извадка участват достатъчно на брой променливи величини, така че дори и след отстраняване на част от тях, останалите ще бъдат достатъчни като брой за изграждане на регресионен модел на взаимовръзка. При някои анализи се предприемат по-либерални подходи, при които се отстранят променливите величини с липсващи стойности между 5 и 10%, като в практиката се е наложило задължителното отстраняване на променливите величини, ако липсващите в тях стойности са в границите между 25 и 30%.

В табл. 3 е представено наличието на липсващи стойности във формираната извадка от данни за това изследване. Техният дял е под 3% за всяка от изследваните променливи в извадката. Следователно тук дори да се избере рестриктивният подход, това не би довело до отстраняване на променливи.

След като са установени наличието и размерът на липсващите стойности във формираната извадка, е необходимо да се пристъпи към тяхното заместване с подходяща стойност. Заместването им зависи от типа на променливите величини. Коректно е категорийните променливи да се заместят с най-често срещаната стойност (модата), но понякога при този вид променливи е по-добре липсващите стойности да се заместят със стойност, която показва отсъствие на изследвания параметър. Например, както вече посочихме, променливата „образование“ допуска следните възможни стойности: 0 - начално, 1 - средно, 2 – средно специално, 3 - полувисше и 4 - висше. От табл. 3 се вижда, че 0,15% от стойностите в нея са липсващи и могат да се заместят с най-често срещаната в

разпределението й стойност, а именно 1 – средно образование. Тъй като не съществува друга степен на образование освен изброените като възможни отговори, то в случаите, когато сред стойностите в тази променлива има липсващи, не би могло да се избере като стойност за заместване друга освен някоя от вече изброените.

По различен начин стои въпросът със заместването на липсващите стойности например в променливата „собствено недвижимо имущество“. Тя приема следните възможни стойности във формираната в изследването извадка от данни: 0 - апартамент/къща; 1 - търговски обект; 2 - земя; 3 - друго; 4 - повече от един апартамент/къща; 5 - апартамент и къща едновременно; 6 - търговски обект и земя едновременно; 7 - повече от един търговски обект; 8 - частична собственост (идеална част).

Таблица 3

Липсващи и отдалечени стойности

Променлива	Отдалечени ст-сти (Outliers)	Екстремни ст-сти (Extremes)	Липсващи ст-сти (брой)	Липсващи ст-сти (%)
Възраст	0	0	0	0.00%
Пол	--	--	0	0.00%
Образование	--	--	112	0.15%
Семейно положение	--	--	93	0.13%
Упражнявана професия	--	--	0	0.00%
Общ трудов стаж в месеци	0	0	87	0.12%
Трудов стаж в години при последния работадател	748	44	176	0.24%
Собствено недвижимо имущество	--	--	86	0.12%
Нетен месечен доход	273	229	0	0.00%
Източник на нетния месечен доход	--	--	93	0.13%
Брой лица в домакинството	767	0	86	0.12%
Брой лица в домакинството, получаващи доход	1613	0	98	0.13%
Брой членове в семейството	947	31	86	0.12%
Брой притежавани МПС	1226	234	102	0.14%
Брой притежавани мобилни телефони	0	0	0	0.00%
От колко години лицето живее на посочения от него адрес	198	0	89	0.12%
Размер на общата кредитна експозиция на клиента	316	406	0	0.00%
Информация за наличие на просрочена част на 1 ден от общата кредитна експозиция на клиента	1272	914	0	0.00%
Информация за наличие на просрочена над 90 дни част от общата експозиция на клиента	204	582	0	0.00%
Обща сума на салдото по депозитни сметки	149	197	0	0.00%
Обща сума на салдото по спестовни сметки	172	160	0	0.00%
Информация дали лицето е поръчител/съдлъжник по съществуващ кредит	593	426	0	0.00%
Отношение на дохода на клиента към бр. лица в домакинството	236	164	86	0.12%
Съотношение на дохода на клиента към бр. лица в домакинството, получаващи доход	266	168	98	0.13%
Съотношение на дохода на клиента към бр. издържани лица	255	175	86	0.12%

От табл. 3 се вижда, че 0,12% от всички стойности в променливата „собствено недвижимо имущество“ са липсващи. В този случай изброените стойности не изчерпват напълно възможностите за отговор на това какво собствено недвижимо имущество притежава даден клиент, когато кандидатства за кредит в банковата система. Ето защо при подобни категорийни променливи е уместно липсващите стойности да се заместят с нова, например „9 – няма“, показваща липсата на собственост. Така манипулирането на липсващите стойности ще остане най-адекватно към реалната ситуация, при която е формирана извадката от данни.

При метрираните променливи отново могат да се изберат два подхода при заместване на липсващите стойности. Първият е заместването им със средната величина, формирана от останалите стойности. Тук за такава е уместно да се избере медианата от съвкупността от останалите стойности. Заместването на липсващите стойности със средноаритметичната стойност не е подходящ вариант, тъй като тя се влияе много силно от екстремните стойности в съвкупността. На този етап от статистическия анализ, отнасящ се до качеството и пълнотата на данните, все още не са предприети стъпки за анализ и коригиране на екстремните стойности в извадката. Ето защо на етапа на анализиране на липсващите стойности и на търсенето на подходящ начин за тяхното заместване не бива да се забравя, че в съвкупността от данни все още съществуват екстремни стойности и те биха повлияли върху формиране на средноаритметичната стойност.

Вторият подход е използването на нулева стойност като индикатор за липса на стойност в измерването. Той подход е по-подходящ при променливи, свързани с бройни единици - брой лица в домакинството; брой лица в домакинството, получаващи доход; брой издържани лица; брой притежавани МПС; брой притежавани мобилни телефони и др.

Дали медианата ще се предпочете като стойност за заместване на липсващите стойности при метрирани променливи, или те ще бъдат заменени с нулеви, е въпрос на избор на всеки анализатор. Основното, което трябва да се вземе предвид в тези ситуации, е това, че колкото по-коригирани са данните, с които се извършва анализът, толкова по-слабо резултатът от него ще отразява реалната ситуация. Всяка намеса в данните чрез коригиране или отстраняване на стойности повлиява върху прогностичната способност на изградения регресионен модел. За да се построи регресионно уравнение с висока познавателна сила, е необходимо корекциите в първоначалните данни от извадката да бъдат минимални.

След корекция и отстраняване на липсващите стойности анализът на извадката с данни за физически лица, ползващи кредити за текущо потребление от банковата система, подлежи на филтриране на отдалечените и екстремните стойности, ако има такива. Отдалечените точки са необикновени и нетипични за наблюдаваната съвкупност. Те могат да се открият с помощта на междуквартилното разстояние (IQ), което е разлика между трети и първи

квартил[3] в съвкупността от данни за всяка от изследваните независими променливи. Ако за стойност X_i от дадена променлива е изпълнено неравенството (1) или (2), тогава стойността се определя като *леко отдалечена (outlier)*:

(1) $X_i < Q1 - 1{,}5 * IQ \quad , \; i = 1, 2, ..., N$

(2) $X_i > Q3 + 1{,}5 * IQ \quad , \; i = 1, 2, ..., N$

Ако за стойност X_i от дадена променлива е изпълнено неравенството (3) или (4), тогава стойността се определя като екстремна (*силно отдалечена) (extreme)*:

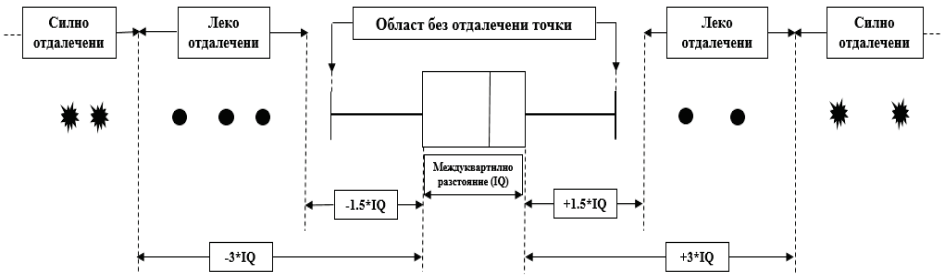
(3) $X_i < Q1 - 3 * IQ \quad , \; i = 1, 2, ..., N$

(4) $X_i > Q3 + 3 * IQ \quad , \; i = 1, 2, ..., N$

Наличието на отдалечени точки може да се визуализира с помощта на „box-plot" графика (фиг. 3).

Фигура 3

„Box-plot"



При откриване на отдалечени точки в извадката от данни е необходимо да се направи анализ за техния произход. Ако те са получени в резултат от грешка при въвеждане на данните и има недвусмислено потвърждение за това, тогава е уместно тези стойности просто да бъдат изтрити. Например отрицателна стойност в променливата „възраст" е ясен знак, че е допусната грешка при генерирането на тази стойност, което изисква подобни стойности да бъдат изтрити от общата извадка (вж. Gujarati, 2004).

Понякога отдалечените точки не могат да бъдат определени като грешки при въвеждане на данните. В тези случаи тяхното отстраняване не е препоръчително, а се преминава към заместването им със стойността, която е равна на израза в дясната страна на неравенства (1), (2), (3) или (4), ако е изпълнено съответно някое от тях.

[3] Квартилите разделят извадката от данни на 4 равни части така, че всяка част съдържа точно 25% от общата честота на данните. Втори квартил съвпада с медианата на разпределението.

От изложеното дотук можем да обобщим, че докато анализът на липсващите стойности се извършва върху цялата извадка, този на отдалечените стойности се прилага само върху част от нея. Идеята е да се изгради регресионно логистично уравнение върху една част от първоначално формираната извадка, наречена тренировъчно множество, а върху останалата да се тества неговото поведение - валидираща подизвадка. Тези две части могат да бъдат подбрани по такъв начин, че съотношението между тях да е 80:20 или 70:30, или 50:50. Изборът зависи от анализатора, но основното правило, което трябва да се спазва, е формирането на двете подизвадки да се изпълни по случаен начин. Тук е избрано съотношение 70:30 съответно между тренировъчното множество и валидиращата подизвадка, при което разпределението на случаите според стойностите на зависимата променлива Y има вида, представен в табл. 4.

Таблица 4

Разпределение на случаите в съотношение 70:30

Извадка от данни	Стойности на незав. променлива Y	Бр.случаи	%
Тренировъчно множество	0 – Добри клиенти	48608	95.23%
	1 – Не добри клиенти	2433	4.77%
Валидираща подизвадка	0 – Добри клиенти	20865	95.37%
	1 – Не добри клиенти	1014	4.63%
Общо		**72920**	**100%**

По-нататък с данните от тренировъчното множество ще анализираме значимостта на независими променливи, които да бъдат включени при изграждане на логистичния регресионен модел. Ще извършим и категоризиране на тези променливи чрез статистическата техника CHAID. Изграденият логистичен регресионен модел ще бъде подложен на тест върху данните от валидиращата подизвадка, резултатите от който ще сравним и ще дадем оценка на качеството на модела.

Избор на независими променливи за изграждане на логистично регресионно уравнение. Избор на регресионен модел

Във втората част от процеса на изграждане на логистично регресионно уравнение за предвиждане на вероятността от попадане в състояние на неизпълнение на кредитоискатели, кандидатстващи за кредит за текущо потребление, се включват етапите по: трансформиране на независимите променливи чрез статистическата техника CHAID; избор на независими променливи, с които да се построи уравнението; прилагане на избраното регресионно уравнение върху валидиращата подизвадка, за да се изпита неговото поведение върху данни, за които не е извършена корекция на отдалечените стойности.

Трансформирането на променливите чрез CHAID метода включва едномерен анализ на взаимодействието на всяка от независимите променливи със зависимата променлива Y. Нека за удобство наречем тази променлива „флаг за неизпълнение“. Чрез метода CHAID се категоризират всички независими променливи, които не са линейно свързани със зависимата. В случаите на непрекъсната независима променлива чрез този метод се създава нова променлива, която е от категориен тип. Последната се получава чрез разделяне на стойностите на първоначалната непрекъсната променлива в интервали (категории) по такъв начин, че случаите, попадащи в отделните категории, да са максимално концентрирани според стойностите на зависимата променлива Y. Дали предложеното чрез CHAID категоризиране е оптимално и статистическо значимо, се установява при изчисляване на хи-квадрат (chi-square) стойност и съответстващата ѝ стойност за статистическа значимост (p-value)[4].

От схемата се вижда, че непрекъснатата променлива „възраст“ е категоризирана в нова, съдържаща четири групи: кредитоискатели на възраст от 27 или по-малко години; кредитоискатели на възраст между 27 и 44 навършени години; кредитоискатели между 44 и 52 навършени години и кредитоискатели над 52 години.

Схема

Формиране на категории при променливата „възраст“
чрез метода CHAID

Зависима променлива Y

Категория	%	брой
0	95.233	48608
1	4.767	2433
Общо	100	51041

Възраст

Adj.P-value=0.000, Chi-square=2068.736, df=3

<=27

Категория	%	брой
0	87.566	4451
1	12.434	632
Общо	9.959	5083

(27;44]

Категория	%	брой
0	89.792	9087
1	10.208	1033
Общо	19.827	10120

(44;5]

Категория	%	брой
0	94.029	5118
1	5.971	325
Общо	10.664	5443

>52

Категория	%	брой
0	98.543	29952
1	1.457	443
Общо	59.55	30395

Категоризирането на променливата „възраст“ чрез метода CHAID може да се определи като статистически значимо, тъй като хи-квадрат стойността е статистически значима (p-value < 0.05).

Категорийните независими променливи също могат да се прекатегоризират чрез метода CHAID. Целта е да се получи оптимално разделяне на променлива в категории по такъв начин, че отново случаите, попадащи в отдел-

[4] Всички статистически тестове и заключения в това изследване се извършват при интервал на доверителност от 95% и съответно ниво от 5% за приемане или отхвърляне на нулевите хипотези в направените статистически тестове.

ните категории, да са максимално концентрирани според стойностите на зависимата променлива Y. След прекатегоризиране чрез CHAID на категорийната променлива „образование“ се получава нова променлива „образование_кат“, която се състои от две категории: категория 1, съдържаща стойности 0, 1 и 2, които съответстват на степените за завършено образование, респ. начално, средно и средно специално, и категория 2, включваща стойности 3 и 4, които съответстват на полувисше и висше образование.

Както беше посочено, при подбор на променливи, които могат да участват в анализа на надеждността на кредитоискателите, е възможно да се създадат нови променливи като резултат от математически операции между наличните до момента. По този начин в хода на представения тук анализ на променливите са създадени следните нови променливи, представляващи съотношение на някои от вече изброените:

- *обща дебитна експозиция на клиента* (като сума от стойностите на двете променливи - обща сума на салдото по депозитни сметки и обща сума на салдото по спестовни сметки);
- *съотношение на дебитна към кредитна експозиция на клиента* (като съотношение на променливите „обща дебитна експозиция на клиента“ към „размер на общата кредитна експозиция на клиента“)
- *флаг за неизпълнение - дали клиентът е поръчител/съдлъжник* по вече съществуващ кредит на лице в банковата система (създадена от променливата „информация дали лицето е поръчител/съдлъжник по съществуващ кредит“).

Същността в подбора на променливите, които могат да участват в изграждането на логистично регресионно уравнение, се изразява в провеждането на няколко статистически теста, които определят значимостта на всяка от независимите променливи относно предвиждането на вероятността за сбъдване на изследваното събитие, т.е. от попадане в състояние на неизпълнение. Тестовете се извършват чрез алгоритъма на „дървото на решенията“, както и чрез прилагане на алгоритъма на невронни мрежи. Към тези статистически тестове е полезно да се добави и проверка за степента на информационната стойност, която носи всяка от категоризираните вече променливи - Information Value Test. С помощта на този тест се анализира тяхната обяснителната сила по отношение на изследваното събитие – вероятност от попадане в състояние на неизпълнение).

Например, когато коефициентът на корелация на Пиърсън между „общ трудов стаж в месеци“ и „флаг за неизпълнение“ (зависимата променлива) има отрицателна стойност (-0,194), това показва слаба, почти липсваща линейна връзка между двете променливи. Ето защо е целесъобразно „общ трудов стаж в месеци“да се категоризира по познатия вече начин чрез метода CHAID. В резултат от неговото прилагане се получава нова променлива от категориен тип - „общ трудов стаж в месеци_кат“. Тя разделя извадката от данни в две категории според значението на стойността, посочена като общ трудов стаж в месеци за

всеки от случаите, както следва: категория 1 „общ трудов стаж в месеци <= 200 месеца“ и категория 2 „общ трудов стаж в месеци > 200 месеца“.

Тестът за определяне на информационната стойност на новата категорийна променлива се състои в калкулиране на следните съотношения между случаите, попадащи в групите и определени от зависимата променлива Y:

- За всяка от категориите в категоризираната променлива се определя броят на 1-те и 0-те, т.е. броят на „добрите“ и „недобрите“ клиенти.
- Калкулира се относителният дял (в %) на 1-те и 0-те във всяка от категориите на категоризираната променлива.
- Изчислява се т.нар. значимост на данните (Weight of Evidence) чрез следната формула:

(5) $$WoE = \ln(pgood / pbad)$$, където:

pgood е делът на случаите, класифицирани като „добри“ в общия брой на „добрите“;
pbad - делът на случаите, класифицирани като „недобри“ в общия брой на „недобрите“

- Накрая информационната стойност се калкулира чрез формулата

(6) $$InformationValue = \sum_{i=1}^{k}[(pgood - pbad)/100 * WoE]$$, където *k* посочва броя на категориите в категорийната променлива.

Информационна стойност, която е по-малка от 0,1, показва ниска обяснителната сила по отношение на изследваното събитие. При изграждане на регресионното уравнение е препоръчително променливи с такава ниска информационна стойност да бъдат отстранени. Информационна стойност, която е поголяма от 0,2-0,3, показва силна обяснителната сила и променливи с такива информационни стойности е много вероятно да участват във финалното регресионно уравнение.

Изборът на независими променливи, с които да бъде изградено логистично регресионно уравнение, може да се направи с помощта на данните в табл. 5, на която са представени резултати за някои от променливите от проведения алгоритъм на „дърво на решенията“, алгоритъма на невронни мрежи и информационните стойности, калкулирани за всяка от тези променливи.

Ако дадена променлива е преминала всеки от посочените в табл.5 тестове, тогава се очаква, че тя има много силна предиктивна способност за определяне на вероятността от поява на събитие на несъстоятелност. Такива променливи е много вероятно да вземат участие в изграждане на логистично регресионно уравнение. Често в практиката се оказва, че изследваните променливи са подложени само на част от тези тестове, а не на всички. Ето защо задача на анализатора е да направи експертна преценка за броя на тестовете, които трябва да „преминат“ анализираните независими променливи.

Таблица 5

Едномерен анализ за значимостта на някои от независимите променливи

Независими променливи	Хи-квадрат статистика			F-статистика			Невронни мрежи
	ст-ст	ст.на своб	вероятност	ст-ст	ст.на своб	вероятност	
Възраст	Chi-square=2125.567	7	0.000	F=2072.137	1; 51039	0.000	0.0028
Възраст_кат	Chi-square=2068.736	3	0.000	Chi-square=2068.736	3	0.000	0.0281
Брой лица в семейството	Chi-square=370.231	2	0.000	F=188.213	1; 51039	0.000	0.0065
Брой лица в семейството_кат	Chi-square=321.200	1	0.000	Chi-square=321.200	1	0.000	0.0124
Брой лица в домакинството, получаващи доход	Chi-square=329.923	2	0.000	F=0.582	1; 51039	1.000	0.0033
Брой лица в домакинството	Chi-square=323.295	2	0.000	F=10.883	1; 51039	0.075	0.0048
Брой лица в домакинството_кат	Chi-square=168.165	1	0.000	Chi-square=168.165	1	0.000	0.0086
Брой притежавани МПС	Chi-square=10.311	1	0.003	F=10.723	1; 51039	0.082	0.0126
Брой притежавани МПС_кат	Chi-square=10.311	1	0.001	Chi-square=10.311	1	0.102	0.0168
Обща дебитна експозиция на клиента	Chi-square=8321.280	5	0.000	F=153.047	1; 51039	0.000	0.0026
Обща дебитна експозиция на клиента_кат	Chi-square=8258.067	1	0.000	Chi-square=8258.067	1	0.000	0.0218

Тук приемаме, че ако независимите променливи удовлетворяват изискванията на поне два от трите теста в табл. 5, тогава те могат да участват в построяване на регресионния модел. Такива променливи са всички с изключение на две – „брой лица в домакинството“ и „брой лица в домакинството, получаващи доход“.

Информационните стойности, калкулирани за независимите променливи, са представени в табл. 6, където участват само променливи от категориен тип.

Таблица 6

Информационна стойност на независимите променливи

Независими променливи	Информационна стойност (Information Value)
Възраст_кат	0.9063
Брой членове в семейството_кат	0.1524
Брой лица в домакинството_кат	0.0606
Брой притежавани МПС_кат	0.0046
Обща дебитна експозиция на клиента_кат	2.2144
Съотношение на дебитна към кредитна експозиция на клиента_кат	0.0662
Флаг, дали клиентът е поръчител/съдлъжник по вече съществуващ кредит на лице в банковата система	0.0444
Пол	0.0639
От колко години лицето живее на посочения от него адрес_кат	0.2092
Трудов стаж в години при последния работодател_кат	0.7549
Образование_кат	0.0350
Семейно положение_кат	0.1787
Собствено недвижимо имущество_кат	0.2028
Размер на общата кредитна експозиция на клиента	0.7506
Брой притежавани мобилни телефони_кат	0.2584
Упражнявана професия_кат	0.8072
Общ трудов стаж в месеци_кат	0.6943
Информация за наличие на проср. част на 1 ден от общата кред. експозиция на клиента_кат	2.9687
Информация за наличие на просрочие над 90 дни част от общата експоз. на клиента_кат	1.1748

Резултатите от анализа на информационните стойности показват, че част от променливите притежават силна предиктивна способност за класифициране на единиците от извадка според случаите, определени от независимата променлива. Променливи, чиято информационна стойност е по-ниска от 0,1, трябва да отпаднат от последващ анализ, както и от участие в изграждане на регресионно уравнение.

Както може да се види от табл. 6, променливите с информационна стойност, по-ниска от 0,1, са:

- брой лица в домакинството_кат;
- брой притежавани МПС_кат - категория 1 „брой притежавани МПС = 0“ и категория 2 „брой притежавани МПС > 0;
- съотношение на дебитна към кредитна експозиция на клиента_кат – категория 1 „съотношение на дебитна към кредитна експозиция на клиента = 0“ и категория 2 „съотношение на дебитна към кредитна експозиция на клиента > 0“;
- пол (1 - мъж; 0 - жена);
- флаг, показващ дали клиентът е поръчител/съдлъжник по вече съществуващ кредит на лице в банковата система (1 - да; 0 - не);
- образование_кат - категория 1 „начално, средно или средно специално/ колеж“ и категория 2 „полувисше и висше“.

Изграждането на логистично регресионно уравнение се извършва с помощта на избраните за статистически значими независими променливи. Като метод за изпълнение на процедурата по построяване на модела е избран вариантът с „обратно връщане“ (Backward Stepwise).

При всяко стартиране на процедурата за изграждане на логистично регресионно уравнение се генерира линейно уравнение от вида:

(7) $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + ... + \beta_k * X_k$, където с $X_1, X_2, ..., X_k$ са обозначени независимите променливи, избрани от статистическите тестове да участват в изграждане на финално уравнение при логистична регресия, а $\beta_0, \beta_1, ..., \beta_k$ са съответните бета-коефициенти за всяка от независимите променливи. При всяко генериране на вариант на уравнение от вида (7) се визуализират резултати от проведени статистически тестове за значимост както на всяка от подбраните независими променливи, така на цялото предложено уравнение. Това са: AIC и BIC критерии за оценяване на доброто изпълнение на модела; хи-квадрат, Likelihood Ratio тест и Pseudo R-Square коефициенти; тест за значимост на бета-коефициентите на всяка от независимите променливи $X_1, X_2, ..., X_k$, взела участие при формирането на даденото уравнение. От особено значение тук е съблюдаването на резултатите от теста за значимост на бета-коефициентите. Ако някой от тях е определен като статистически незначим, тогава съответната му независима променлива трябва да се отстрани от анализа в създаването на последващи регресионни модели. Много често в

практиката се оказва, че независимите променливи имат висока предиктивна способност (според проведения едномерен анализ) по отношение на вероятността на изследваното събитие, но в комбинация с други независими променливи тази способност на някои от тях се влошава и променливата става статистически незначима. Това се дължи основно на общото действие на всички променливи, взети заедно, при определяне на най-вероятната стойност според зависимата променлива Y, която може да се случи като възможен резултат на основата на всички наблюдения във формираната извадка от данни.

След множество изпълнения на статистическата процедура за формиране на логистично регресионно уравнение от вида (7) като краен резултат е избрано следното (табл. 7):

Таблица 7

Уравнение за изграждане на логистична регресия

Променливи, участващи в логистичното регресионно уравнение								
Променливи	бета - коефициент	Станд. Грешка	Wald Статистика	Степени на свобода (df)	Коефициент на значимост (Sig.)	Exp (бета-коефициент)	95.0% Интервал на доверителност за Exp(бета- коефициент)	
							Долна граница	Горна граница
Константа	-3.274	0.14	548.061	1	0			
Обща дебитна експозиция на клиента	-0.02554	0.001	428.156	1	0	0.975	0.972	0.977
Размер на общата кредитна експозиция на клиента	0.00004692	0	199.747	1	0	1	1	1
Семейно положение_кат=1	-0.3002	0.126	5.644	1	0.018	0.741	0.578	0.949
Семейно положение_кат=2	0(b)	.	.	0	.	.	.	.
Брой членове в семейството_кат=1	0.3754	0.122	9.489	1	0.002	1.456	1.146	1.848
Брой членове в семейството_кат=2	0(b)	.	.	0	.	.	.	.
Трудов стаж в години при последния работадател_kat=1	0.8825	0.067	173.615	1	0	2.417	2.12	2.756
Трудов стаж в години при последния работадател_kat=2	0.3382	0.072	21.966	1	0	1.402	1.217	1.616
Трудов стаж в години при последния работадател_kat=3	0(b)	.	.	0	.	.	.	.
Общ трудов стаж в месеци_кат=1	0.3415	0.057	36.054	1	0	1.407	1.259	1.573
Общ трудов стаж в месеци_кат=2	0(b)	.	.	0	.	.	.	.
Упражняване професия_кат=1	-0.7451	0.059	160.35	1	0	0.475	0.423	0.533
Упражняване професия_кат=2	0(b)	.	.	0	.	.	.	.
Отношение на дохода на клиента към размер на общата кредитна експозиция на клиента_кат=1	0.883	0.061	183.634	1	0	2.3	2.039	2.595
Отношение на дохода на клиента към размер на общата кредитна експозиция на клиента_кат=2	0(b)	.	.	0	.	.	.	.

В избраното уравнение участват 8 независими променливи, от които 2 са непрекъснати, а останалите 6 са категорийни променливи, получени чрез категоризиране на първоначалните независими променливи, а именно:

- обща дебитна експозиция на клиента;
- размер на общата кредитна експозиция на клиента;
- семейно положение, категоризирана в две групи - категория 1: „женен/омъжена“ и категория 2 „неженен/неомъжена; разведен/разведена; вдовец/вдовица“;

• брой членове в семейството, категоризирана в две групи - категория 1 „брой членове в семейството <= 1“ и категория 2 „брой членове в семейството > 1“;

• трудов стаж в години при последния работодател, категоризирана в три групи - категория 1 „трудов стаж в години при последния работодател <= 1“; категория 2 „трудов стаж в години при последния работодател > 1“ и „трудов стаж в години при последния работодател <= 4“ и категория 3 „трудов стаж в години при последния работодател > 4“;

• общ трудов стаж в месеци, категоризирана в две групи - категория 1 „общ трудов стаж в месеци <= 200“ и категория 2 „общ трудов стаж в месеци > 200“;

• упражнявана професия, прекатегоризирана в две групи - категория 1 „мениджъри, специалисти или пенсионери“ и категория 2 „квалифицирани или неквалифицирани работници; студенти или лица, упражняващи свободни професии“;

• отношение на дохода на клиента към размера на общата му кредитна експозиция, категоризирана в две групи - категория 1 „отношение на дохода на клиента към размера на общата му кредитна експозиция <= 0,02799“ и категория 2 „отношение на дохода на клиента към размера на общата му кредитна експозиция > 0,02799“.

Всяка от независимите променливи в крайното уравнение е статистически значима (sig. < 0.05) и притежава логически коректна стойност за бета-коефициента си. Ако бета-коефициентът е с положителен знак, това означава, че съответната независима променлива има положителна (права) връзка със събитието за риск, т.е. колкото е по-висок бета-коефициентът, толкова по-голяма е вероятността от попадане в състояние на неизпълнение.

Ако означим със Z уравнението от табл. 7, а именно:

-3.274
-0,02554 * Обща дебитна експозиция на клиента +
0,00004692 * Размер на общата кредитна експозиция на клиента +
-0,3002 * Семейно положение_кат=1 +
0 * Семейно положение_кат=2 +
0,3754 * Брой членове в семейството_кат=1 +
0 * Брой членове в семейството_кат=2 +
0,8825 * Трудов стаж в години при последния работодател_кат =1 +
0,3382 * Трудов стаж в години при последния работодател_кат =2 +
0 * Трудов стаж в години при последния работодател_кат =3 +
0,3415 * Общ трудов стаж в месеци_кат =1 +
0 * Общ трудов стаж в месеци_кат =2 +
-0,7451 * Упражнявана професия _кат =1 +
0 * Упражнявана професия _кат =2 +
0,833 * Отношение на дохода на клиента към размера на общата кредитна експозиция на клиента _кат =1 +
0 * Отношение на дохода на клиента към размера на общата кредитна експозиция на клиента _кат =2

тогава вероятността от попадане в състояние на неизпълнение (PD) може да се изчисли по следният начин:

(8) $$PD = \frac{1}{1+\exp(-Z)}.$$

Уравнение (8) представлява уравнение на логистична регресия, чиито резултативни стойности са разположени в интервала от 1 до 100%.

Преди да се оценят способността и поведението на уравнението на логистична регресия, е необходимо да се направи тест за мултиколинеарност на финалните независимите променливи, участващи в него. Мултиколинеарността се наблюдава, когато независимите променливи в уравнението са линейно свързани, т.е. съществува корелационна връзка между тях. За откриване на наличие на мултиколинеарност се използва следната формула, изчисляваща Variance Interference Factor (VIF):

(9) $$VIF = \frac{1}{(\text{Tolerance})} = \frac{1}{1-R^2}$$, където

R^2 е коефициент на определение между независимите променливи. Като практическо правило се приема, че ако стойностите на VIF са по-големи от 10, тогава съществува силна корелация между тях. Ако се установи наличие на мултиколинеарност, тогава е необходимо да се отстранят променливите, които корелират помежду си.

Променливите, участващи в избраното логистично регресионно уравнение, са изследвани за наличие на мултиколинеарност (табл. 8). Вижда се, че стойностите на VIF са близки до 1, което показва, че не съществува мултиколинеарност между тези променливи.

Таблица 8

Стойности на коефициента Variance Interference Factor

Променливи	Нестандартизирани коефициенти		t - стат	значимост	95,0% Доверителен интервал за бета-коефициента		Collinearity Statistics	
	бета-коеф.	Станд. Грешка			Долна граница	Горна граница	Tolerance	VIF
Константа	0.127	0.003	40.206	0	0.121	0.133		
Обща дебитна експозиция на клиента	-0.0000292	0	-13.694	0	0	0	0.987	**1.013**
Размер на общата кредитна експозиция на клиента	0.00000505	0	31.674	0	0	0	0.786	**1.273**
Брой членове в семейството	-0.028	0.002	-18.568	0	-0.031	-0.025	0.952	**1.051**
Трудов стаж в години при последния работадател	-0.002	0	-13.256	0	-0.002	-0.002	0.812	**1.232**
Общ трудов стаж в месеци	0	0	-24.882	0	0	0	0.782	**1.278**
Отношение на дохода на клиента към размер на общата кредитна експозиция на клиент	0.003	0.003	1.152	0.249	-0.002	0.008	0.863	**1.159**

Тези променливи притежават стилистически значимо оптимално деление според метода CHAID, което позволява прилагането им при оценяване на надеждността на кредитоискателите.

Оценяване и анализ на представянето на избрания регресионен модел

За да се приеме, че регресионното уравнение (8) е приложимо като скоринг-модел, е необходимо да се изследва и неговото поведение. За целта се анализира надеждността и способността на избрания логистичен модел да класифицира коректно единиците в извадката според групите, формирани от зависимата променлива. Прогностичната способност на модела се проверява чрез калкулиране на GINI коефициент както върху данните от тренировъчното множество, така и върху валидиращата извадка.

Стойността на GINI коефициента се калкулира по следната формула:

(10) $$GINI = \frac{A}{A+B}$$, където

с *A* е означена площта под кривата, показваща кумулативния процент на „недобрите“ случаи за всеки от персентилите. Частта, означена с *B*, допълва площта *A* до квадрат със страна 100%.

Прилагайки формула (10) за тренировъчното множество и валидиращата извадка, се получават стойности на GINI коефициента съответно 76,69 и 76,86%. GINI коефициентът, калкулиран върху цялата извадка (тренировъчно множество + валидираща извадка), има стойност 75,66%, която показва, че 75,66% от всички случаи, обозначени с 1 („недобри) според зависимата променлива, се класифицират коректно от избраното логистично регресионно уравнение. От така получените стойности на GINI коефициента може да се направи заключението, че регресионният модел има добра предиктивна способност при класифициране на случаите според значенията на зависимата променлива.

Като допълнение на GINI коефициента и за постигане на по-голяма увереност в пригодността на избрания регресионен модел може да се тества неговата категоризираща сила чрез класификационния тест на Колмогоров-Смирнов:

(11) Колмогоров-Смирнов статистика = Максимална стойност [кумулативен %(добри) - кумулативен %(недобри)]

Чрез този тест се измерва най-голямата дистанция между кумулативния процент на случаите, класифицирани като „добри“, и този на „недобрите“. Резултатите от теста, проведен с данните от изследваната тук извадка, са представени в табл. 9

Максималната стойност на класификационния тест на Колмогоров-Смирнов е 60,87%, която се достига при вероятност от 0,05, където се намират 79,47% от случаите, класифицирани като „добри“, и 18,60% - като „недобри“.

Таблица 9

Класификационен тест на Колмогоров-Смирнов

Вероятност	Зависима променлива (Y)	Добри (флаг)	Не добри (флаг)	Общо	Кумулативен % (Добри)	Кумулативен % (Не добри)	Колмогоров - Смирнов статистика
0.00000000	0	1	0	1	0.001%	0.000%	0.001%
0.00000000	0	1	0	1	0.003%	0.000%	0.003%
0.00000000	0	1	0	1	0.004%	0.000%	0.004%
0.00000000	0	1	0	1	0.006%	0.000%	0.006%
0.00000000	0	1	0	1	0.007%	0.000%	0.007%
0.00000000	0	1	0	1	0.009%	0.000%	0.009%
…	…	…	…	…	…	…	…
0.05254774	0	1	0	1	79.47%	18.60%	60.870%
0.05254849	0	1	0	1	79.47%	18.60%	60.871%
0.05254971	**0**	**1**	**0**	**1**	**79.47%**	**18.60%**	**60.872%**
0.05255637	1	0	1	1	79.47%	18.62%	60.843%
0.05256336	0	1	0	1	79.47%	18.62%	60.845%
…	…	…	…	…	…	…	…
0.99985254	0	1	0	1	99.99%	99.97%	2.300%
0.99997838	0	1	0	1	100.00%	100.00%	0.000%

Проведеният анализ на значимостта на променливите величини и получените резултати за GINI коефициента и класификационния тест на Колмогоров-Смирнов показват, че представеният логистичен регресионен модел може да се използва като скоринг-модел за определяне на надеждността на кредитоискателите в процеса на кандидатстване за кредит.

Скоринг-моделите се основават на исторически данни, но техните резултати се отнасят за бъдещи периоди. Ето защо дори без наличието на промени в икономическата и финансовата среда е необходимо този вид модели да се преоценяват периодично и да се заменят с нови, изградени върху по-актуални данни.

*

От получения логистичен регресионен модел може да се направи изводът, че оценяването на кредитния риск при разрешаване на кредити за текущо потребление на граждани може да се определи чрез 8 от общо 26 първоначално избрани независими променливи. Тези променливи са статистически значимите индикатори от така формираната извадка, с помощта на които може да се определи равнището на риск за всеки кредитоискател в процеса на разрешаване на кредити. Влиянието на всяка от променливите в логистичното регресионно уравнение върху равнището на риск се определя от съответния й бета-коефициент. Колкото по-голяма е неговата стойност, толкова по-висока е вероятността кредитоискателят да попадне в състояние на неизпълнение.

Две от независимите променливи в логистичното регресионно уравнение са непрекъснати – „обща дебитна експозиция на клиента" и „размер на общата кредитна експозиция на клиента". Първата от тях има отрицателен бета-коефициент, равен на -0,02554, което показва, че при нарастване на стойността ѝ с единица вероятността кредитоискателя да попадне в състояние на неизпълнение ще намалее с 2,554% при равни други условия, т.е. по-голям размер на наличните дебитни средства на клиента в момента на кандидатстване за кредит води до по-ниско равнище на риск. Втората непрекъсната променлива има положителен бета-коефициент, равен на 0,00004692, което показва, че при увеличаване на стойността ѝ с единица вероятността кредитоискателят да попадне в състояние на неизпълнение също се повишава с 0,004692% при равни други условия, т.е. по-големият размер на кредитната експозиция на клиента, съществуваща в момента на кандидатстване за кредит, води до по-високо равнище на риск.

Останалите 6 независими променливи в логистичното регресионно уравнение са категорийни като всяка от категориите има съответен бета-коефициент. Променливата „семейно положение", участваща в уравнението, има две категории, като бета-коефициентът на първата е отрицателен и е равен на -0,3002. Това означава, че при равни други условия семейните лица носят по-малък риск, отколкото тези, които не са семейни.

Категоризираната променливата „брой членове в семейството" има две категории. Първата от тях има положителен бета-коефициент, равен на 0,3754, което показва, че ако при равни други условия кредитоискателят е единственото лице в семейството, неговото равнище на риск ще бъде по-високо в сравнение с такъв, който живее в семейство с двама или трима членове.

Като резултат от обяснителната сила на всяка от осемте независими променливи, участващи в регресионното уравнение, може да се обобщи, че според статистическия анализ, извършен с данните от извадката, клиентът - носител на най-ниска степен на риск от попадане в неизпълнение на бъдещите си кредитни задължения, ще има следния рисков профил:

- семеен, упражняващ професия като мениджър или специалист, или е пенсионер;
- има възможно най-висока дебитна експозиция и няма кредитни задължения в момента на кандидатстване;
- работи повече от 4 години за сегашния си работодател и има общ трудов стаж над 200 месеца.

Разнообразието в рисковия профил на лицата, кандидатстващи за кредит, е голямо. Много често част от клиентите - носители на висок риск според бета-коефициента на дадена категория на независима променлива от логистичното регресионно уравнение, попадат в категория, носеща им нисък риск според бета-коефициента на категория от друга независима променлива. По този начин кредитоискатели с различен рисков профил според независимите променливи в регресионното уравнение могат да се окажат носители на един и същи риск

от попадане в състояние на неизпълнение. Това налага да се анализира и оцени кредитният риск и след разрешаване на кредита до неговото пълно погасяване от страна на длъжника.

Използването на скоринг-модели при кандидатстване за кредит е само една част от управлението на кредитния риск, която позволява банките да разграничат надеждните от ненадеждните кредитоискатели. Модерната банкова практика изисква прилагането на подходящ инструментариум от статистически и експертни техники за оценяване на вече разрешените кредити. Това разкрива широка област за бъдещи анализи и оценки на поведението на кредитоискателите.

Използвана литература:

Draper, N. R. and H. Smith (1981). Applied Regression Analysis. 3rd ed. A Wiley-Interscience Publication: John Wiley & Sons Inc.

Feschijan, D. (2008). Analysis of the creditworthiness of bank loan applications. Fasta Universitats.

Finlay, S. (2010). Credit Scoring, Response Modelling and Insurance Rating - A practical guide to forecasting consumer behaviour. Palgrave Macmillan.

Gordon, L. (2013). Using classification and Regression Trees (CART) in SAS. University of Kentucky.

Gujarati, D. (2004). Basic Econometrics, Fourth Edition. The McGraw Hill Companies.

Hosmer, D. W. and S. Lemeshow (2000). Applied Logistic Regression, 2nd ed. New York: John Wiley & Sons Inc.

Lando, D. (2004). Credit Risk Modeling: Theory and Applications. Princeton: Princeton University Press.

Lund, B. and D. Brotherton (2013). Information Value Statistic. Detroit, MI, Paper AA-14-2013

Loh Wei-Yin (2011). Classification and regression trees. Department of Statistics, University of Wisconsin – Madison, USA: John Wiley & Sons Inc.

Madan, D. and H. Unal (1998). Pricing the risks of default. - Review of Derivatives Research, 2, p. 121-160.

15.IV.2015 г.