УДК 004.942

В.Г. Гуськова, аспірантка,
П.І. Бідюк, д-р техн. наук., професор,
1Національний технічний університет України «Київський політехнічний
інститут імені Ігоря Сікорского»
guskovavera2009@gmail.com

Аналіз кредитоспроможності позичальників кредитів за допомогою логістичної регресії

В роботі запропоновано підхід на основі логістичної регресії для аналізу кредитоспроможності позичальників з використанням пакета RStudio. Підхід засновано на виборі, аналізі та застосуванні фактичних вихідних даних для оцінки кредитних ризиків в майбутньому. Використання площі під ROC-кривої та її значення дозволять знайти кращу повну та скорочену моделі, які використовуються в ході роботи. Проведено порівняльний аналіз отриманих моделей та статистичних критеріїв і наведено приклад роботи даного підходу з використанням фінансових даних.

Ключові слова: логістична регресія, аналіз кредитоспроможності, RStudi, кредитний ризик, ROC-крива, статистичні критерії.

DOI: 10.31474/1996-1588-2017-2-25-54-59

Вступ

Сьогодні економіка, фінансовий та банківський сектори перебувають у достатньо важкому стані, що значно впливає на погіршення фінансових результатів діяльності банків, скорочення обсягів депозитів, надання кредитів та значного падіння фінансової стійкості банківських установ. Як наслідок, це все призводить до того, що велика кількість банків стають неплатоспроможними. Цим підтверджується необхідність розробки актуальних заходів, спрямованих на подолання кризи банківського сектору. Відсутність ретельного аналізу стану позичальника та видача фінансів ненадійним особам також сприяли виникненню великих заборгованостей і неповернень кредитів. Це можна пов'язати з тим, що банки не користуються сучасними методиками оцінювання клієнтів або не вдосконалюють та не адаптують їх.

Нові підходи та методи оцінювання ризиків дозволили б банкам відновити кредитування та значно зменшити витрати. Таким чином, необхідно запропонувати підхід, який зробить неможливими відкидання факторів впливу на основну змінну. На етапі збору статистичних даних завчасно невідомі ті фактори, які спричиняють будь-який вплив на результат, через це фінансовими установами збирається вся інформація на основі чого будується модель та визначаться вплив факторів на результат. На цьому етапі аналізуються отримані дані і виявляються суттєві фактори [1].

Застосування методу бінарної логістичної регресії дозволяє досліджувати залежність дихотомічних змінних від незалежних змінних. Залежна змінна приймає два значення (в кредитному скорингу – дефолт/відсутність дефолту) і має біноміальний розподіл.

Постановка задачі

Метою даної статті є розробка кращої моделі для оцінювання платіжної спроможності позичальників кредитів і вибір перспективної моделі для оцінювання ризиків у майбутньому. Для знаходження кращої моделі необхідно застосувати такі показники адекватності моделі, як площа під ROC-кривою та індекс Акайке. Для виконання порівняльного аналізу методів оцінювання побудуємо повні та скорочені моделі та порівняємо значення коефіцієнтів.

Означення логістичної моделі

Логістична регресія – статистична регресійний модель, яку застосовують у випадку, коли залежна змінна є категорійною, тобто може набувати тільки двох значень (0 або 1). Нехай є деяка випадкова величина Y, що може набувати лише двох значень, які, як правило, позначаються цифрами 0 і 1. Нехай ця величина залежить від деякої множини пояснювальних змінних $x = (1, x_1, ..., x_n)^T$ [2, 3].

Залежність Y від x1, x2,..., xn можна визначити завдяки введенню додаткової змінної y*, де

$$y = \theta^T x = \theta_0 + \theta_0 x_1 + ... + \theta_n x_n + \varepsilon \quad (1)$$

тоді:

$$y = \begin{cases} 0, y \le 0 \\ 1, y \ge 0 \end{cases} \quad (2)$$

При побудові логістичної моделі стохастичний доданок ε вважається випадковою величиною з логістичним розподілом ймовірностей. Відповідним конкретним значенням змінних x = x1, x2,..., xn відповідає значення y* і ймовірність того, що y = 1 дорівнює:

$$p(y = 1) = p(y \ge 0) = p(\theta^T x + \varepsilon \ge 0) =$$
$$p(\varepsilon \ge -\theta^T x) = p(\varepsilon \le \theta^T x) = \Lambda(\theta^T x) \quad (3)$$

Логіт-модель задовольняє такій умові:

$$\ln\frac{(pp|X)}{1-p(1|X)} = \ln\frac{(pp|X)}{1-p(1|X)} = b_0 + b_1x_1 + \ldots + b_jx_j \quad (4)$$

Оцінювання параметрів моделі

Оцінювання параметрів $\theta_0 + \theta_1 + \ldots + \theta_n$ виконується на основі вибірки $(x^{(1)}),(Y^{(1)}),\ldots,(x^{(m)}),(Y^{(m)})$, де $x^{(i)} \in R^n$ – вектор значень незалежних змінних, а $Y^{(i)} \in \{0,1\}$ – відповідне їм значення Y, як правило здійснюється за допомогою методу максимальної правдоподібності, згідно з яким вибираються параметри θ, що максимізують значення функції правдоподібності на вибірці:

$$\theta = \operatorname{argmax}_\theta L(\theta) = \operatorname{argmax}_\theta \prod_{i=1}^{m} P\text{r}У = У^{(i)} | x = x^{(i)} \quad (5)$$

Для максимізації цієї функції може бути застосований, наприклад, метод градієнтного спуску, метод Ньютона чи стохастичний градієнтний спуск.

В якості методів оцінювання прогностичної здатності моделі будемо використовувати ROC-аналіз. ROC-крива – графік, що дозволяє оцінити якість бінарної класифікації. Вона також відома як крива помилок. Кількісну інтерпретацію ROC дає показник AUC (англ. Area under ROC curve, площа під ROC-кривою) – площа, обмежена ROC-кривою и віссю частки помилок позитивних класифікацій. Чим вищий показник AUC, тим якісніше функціонує класифікатор; при цьому значення 0,5 демонструє непридатність обраного методу класифікації [4].

Реалізація логістичної регресії

Розглянемо конкретний приклад розробки скорингової моделі за допомогою методу логістичної регресії в системі R. Модель повинна дати прогноз ризиків по клієнтах, які планують взяти кредит в банку, на основі історичних даних [5].

Історична вибірка – дані про 10000 позичальників, які вже скористалися кредитом (табл. 1).

Незалежні змінні:

– вік позичальника;

– освіта позичальника;

– сімейний стан позичальника;

– термін проживання позичальника за останньою адресою, кількість років;

– щомісячний особистий дохід позичальника;

– боргові зобов'язання позичальника від доходу;

– борг позичальника по кредитній карті банку;

– інші боргові зобов'язання позичальника.

Залежна змінна – наявність/відсутність у клієнта боргів по раніше взятих кредитах (default, payment.next.month).

За кожною змінною отримаємо статистичну інформацію (таблиця 2) у програмі R та значення мінімуму, максимуму, середнього і квартилей: першого, другого (медіани) і третини.

Таблиця 1. Історичні дані позичальника

Age	Education	Marriage	Adress	Limit	Bill	AMT 1	AMT 2	N ext pay
24	2	1	3	20000	3913	0	689	1
26	2	2	5	12000	2682	0	1000	1
...	...							
57	2	1	5	50000	8617	2000	36681	0

Таблиця 2. Статистична інформація за змінними

Age		Education		Adress		Bill	
Min	21.0	Min	0.0	Min	1.0	Min	-15308
1st Qu	28.0	1st Qu	1.0	1st Qu	3.0	1st Qu	3378
Median	34.0	Median	2.0	Median	5.0	Median	21958
Mean	35.39	Mean	1.81	Mean	5.49	Mean	49856
3rd Qu	41.0	3rd Qu	2.0	3rd Qu	8.0	3rd Qu	64601
Marriage		AMT1		AMT2		Next pay	
Min	0.0	Min	0	Min	0	Min	0.0
1st Qu	1.0	1st Qu	963	1st Qu	650	1st Qu	0.0
Median	2.0	Median	2103	Median	2000	Median	0.0
Mean	1.57	Mean	5619	Mean	5783	Mean	0.2259
3rd Qu	2.0	3rd Qu	5006	3rd Qu	5000	3rd Qu	0.0

Відмітимо, що змінна education є не кількісною, а категоріальною, тому для подальшого аналізу її треба замінити на штучні змінні, які будуть відповідати за кожен із рівнів змінної education.

```
for(i in 1:6) {
        assign(paste("EDUCATION",".",i,sep=""),
        rep(0, length(DataModel$EDUCATION)))
}
>EDUCATION.1[DataModel$EDUCATION==1]<- 1
> EDUCATION.2[DataModel$EDUCATION==2]<- 2
> EDUCATION.3[DataModel$EDUCATION==3]<- 3
> EDUCATION.4[DataModel$EDUCATION==4]<- 4
> EDUCATION.5[DataModel$EDUCATION==5]<- 5
> EDUCATION.6[DataModel$EDUCATION==6]<- 6
```

Для кожної змінної отримаємо рівняння для переходу на кількісне значення.

Побудова регресійної моделі

Побудуємо модель логістичної регресії безпосередньо за вихідними змінними і змінними, перетвореними у взаємонезалежні. Цей крок необхідно зробити для того, щоб виконати одне з найважливіших умов методу логістичної регресії – відсутність мультиколінеарності (кореляції між предикторами).

Спочатку побудуємо повну модель логістичної регресії за всіма вихідними предикторами:

```
> model <- glm(default.payment.next.month ~ ADRESS +
EDUCATION.0 + EDUCATION.1 + EDUCATION.2 +
EDUCATION.3 + EDUCATION.4 + EDUCATION.5
+EDUCATION.6 + MARRIAGE + LIMIT_BAL +
PAY_AMT1 + PAY_AMT2 + BILL_AMT1,
family=binomial(logit),DataModel)
```

Виведемо статистичну інформацію про побудовану модель (таблиця 3):

```
> Summary (DataModel)
```

Таблиця 3. Статистичні критерії за змінними

Min	1Q	Median	3Q	Max
-0.9655	-0.7812	-0.6767	-0.2945	4.3332

Значення критерію Акайке дорівнює AIC: 10379.

Виводиться така інформація:

– дані про залишки;

– регресивні коефіцієнти, їх стандартні помилки і рівні значущості;

– інформаційний критерій Акайке (AIC), за якими визначають якість моделі [5, 6].

При додаванні змінних в модель якість підгонки зазвичай збільшується. На практиці точна специфікація моделі невідома, неможливо заздалегідь визначити, які предиктори включати, а які ні. Тому користуються покроковою регресією. В результаті вибираються тільки ті предиктори, які дозволять побудувати модель, оптимальну по інформаційному критерієм Акайке (AIC). Даний критерій враховує впливи на якість підгонки моделі, дозволяючи вибрати з безлічі моделей найкращу.

Оскільки мова йде про скорочення числа предикторів, побудовану в результаті модель ще називають скороченою.

Побудуємо скорочену модель логістичної регресії, оптимальну за критерієм AIC:

```
> model.f <- step(model)
Start:  AIC=10379.34
default.payment.next.month ~ AGE + MARRIAGE +
ADRESS + LIMIT_BAL +
  BILL_AMT1 + PAY_AMT1 + PAY_AMT2 +
EDUCATION.1 + EDUCATION.2 +
  EDUCATION.3 + EDUCATION.4 + EDUCATION.5 +
EDUCATION,6
```

Для цього необхідно використовувати процедуру PCA (principal component analysis – метод головних компонент), яка замінить набір предикторов на еквівалентний їм множину взаємно незалежних показників (компонент).

Трансформуємо дані за допомогою PCA, перетворивши їх в набір взаємно незалежних процесів:

```
> data.pca <- prcomp(cbind(DataModel[,-c(9,2)]),
center=FALSE, scale,=FALSE)
```

Додаємо до трансформованих предикторів залежну змінну:

```
> data.prcomp <- data.frame(data.frame(data.pca$x),
default.payment.next.month=DataModel$default.payment.n
ext.month)
```

Аналіз принципових компонент замінює вихідну множину з семи предикторів на множину, що складається з такого ж числа лінійних комбінацій вихідних змінних (PC1, PC2 ..., PC7).

Тепер будуємо повну модель логістичної регресії з використанням трансформованих предикторів:

```
> model.prcomp <- glm (default.payment.next.month ~
PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7,
family=binomial(logit), data,prcomp)
```

Виводимо інформацію про побудовану модель:

```
> summary(model,prcomp)
```

Таблиця 4. Статистичні критерії за змінними

Min	1Q	Median	3Q	Max
-0.9578	-0.7776	-0.6776	-0.3173	4.3256

Значення критерію Акайке дорівнює AIC: 10396.

У таблиці 4 наведені статистичні критерії при використанні трансформованих предикторів та у таблиці 5 наведено результати прогнозу моделі за початковою вибіркою.

Таблиця 5. Результати прогнозу залежної змінної

1	2	3	4	5	6	7	8
0.32	0.23	0.25	0.31	0.17	0.27	0.02	0.25
9	10	11	12	13	14	...	10000
0.25	0.28	0.20	0.10	0.08	0.27	...	0.19

Таким чином, по кожному позичальнику визначено ймовірність повернення ним кредиту.

ROC-аналіз регресійної моделі

В якості методів оцінки прогностичної здатності моделі ми використовували ROC-аналіз. Моделлю є будь-який бінарний класифікатор (в нашому прикладі це логіт-модель). Позитивним результатом буде наявність дефолту у позичальника, а негативним результатом – відсутність дефолту.

На графіку ROC-кривої по осі ординат відкладається чутливість (істинно позитивні приклади), по осі абсцис – або специфічність (істинно негативні приклади), або 1 мінус специфічність (помилково позитивні приклади) [7].

Чим крива більш вигнута (ближче до верхнього лівого кута), тим вище здатність моделі. Та навпаки, чим ближче вона розташована до діагональної прямої, тим менш ефективна модель.

Розраховуємо ROC-криву для повної моделі:

```
> roc.model <- roc(data$default, predict(model,
type="response"), ci=TRUE)
```

Розраховуємо ROC-криву для скороченої моделі:

```
> roc.model.f <- roc(data$default, predict(model.f,
type="response"), ci=TRUE)
```

Розраховуємо ROC-криву для повної моделі за даними PCA:

```
> roc.model.prcomp <- roc(data$default,
predict(model.prcomp, type="response"), ci=TRUE)
```

Розраховуємо ROC-криву для скороченої моделі за даними PCA:

```
> roc.model.prcomp.f <- roc(data$default,
predict(model.prcomp.f, type="response"), ci=TRUE)
```

Побудуємо графік ROC-кривої для повної моделі (рисунок 1): > plot.roc(roc.model)

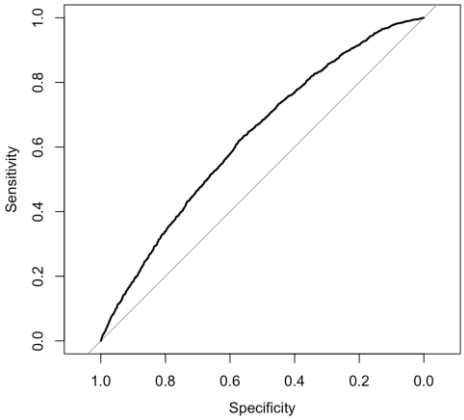

Рисунок 1 – ROC-крива для повної моделі

Побудуємо графік ROC-кривої для повної моделі за даними PCA (рисунок 2):

```
> plot.roc(roc.model.prcomp, add=TRUE,
col="red")
```

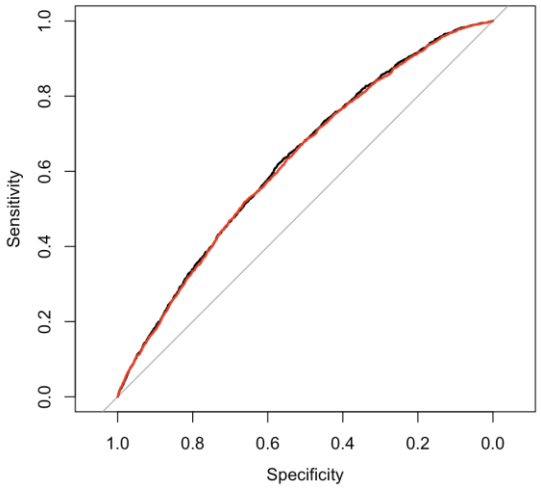

Рисунок 2 – ROC-крива для повної моделі за методом головних компонент

Побудуємо графік ROC-кривої для скороченої моделі за даними PCA (рисунок 3). З рисунків бачимо, що значення кривої та площі візуально не сильно відрізняються. Для того, щоб знати яку саме модель нам використовувати, ми повинні розрахувати чисельні значення площі для кожного графіку та на основі отриманого результати обрати кращу модель.

```
> plot,roc(roc.model.prcomp.f, add=TRUE, col=«green»)
```

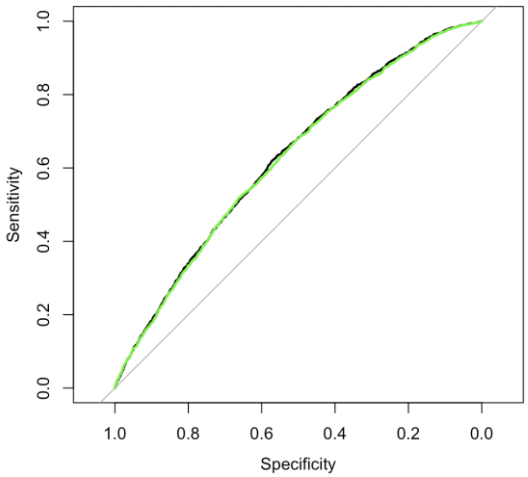

Рисунок 3 – ROC-крива для скороченої моделі

Порівняльний аналіз ROC-кривих для моделей

Запустимо непараметрический тест Делонга для аналізу ROC-кривих. Порівнюємо повну модель і редуковану модель, які були побудовані за вихідними даними [7]:

```
>roc.test(roc.model, roc.model.f,  method="delong")
data:                roc.model          roc.model.f
                     Z = 0.51887        p-value = 0.6039
alternative hypothesis: true difference in AUC is not equal
```

to 0
sample estimates: AUC of roc1 AUC of roc2
0.62923 0.62881

Значення площі під кривою становить більше ніж 0.5, це означає, що ми можемо використовувати дві моделі, але кращий результат буде при роботі з повною моделлю.

Порівнюємо редуковану модель за вихідними даними і редуковану модель за трансформованими даними:

> roc.test(roc.model.f, roc.model.prcomp.f, method=«delong")
data: roc.model roc.model.f
Z = 3.2787 p-value=0.001043
alternative hypothesis: true difference in AUC is not equal to 0
sample estimates: AUC of roc1 AUC of roc2
0.62881 0.62563

За результатами тесту Делонга та порівнянню статистичних критеріїв моделі найкращою є повна модель за вихідними даними. Треба зазначити, що доцільно використання і інші отримані моделі логістічної регресії при оцінюванні кредитоспроможності позичальників.

Висновки

Ретельний аналіз характеристик клієнтів дає можливість банкам набагато ефективніше та раціональніше формувати свої кредитні програми. На основі логістичних моделей були отримані високі значення AUC = 0.6292. Результати показують, що в банківським закладам потрібно використовувати скорингові моделі, таким чином зменшуючи обсяги втрат банків від неповернення кредитів.

У подальших дослідженнях доцільно розробити і застосувати комбінований підхід для оцінювання ймовірності дефолту позичальника на основі мереж Байєса та нейронних мереж.

Список літератури

1.Бідюк П.І. Моделі оцінки ризиків кредитування фізичних осіб / Бідюк П.І., Матрос Є.О. // Кібернетика та обчислювальна техніка. — 2007. — No 153. — С. 87—95.
2. Терентьев А.Н. Сравнение методов интеллектуального анализа данных при оценивании кредитоспособности физических лиц / Терентьев А.Н., Бидюк П.И., Миронова А.В., Медин Н.Ю. // Пробл. управления и информатики. — К.: ИКИ НАНУ-НКАУ, 2009. — No 5. — С. 141—149.
3. Кузнєцова Н.В. Системний підхід до аналізу кредитних ризиків з використанням мереж Байєса / Кузнєцова Н.В., Бідюк П.І. //Наукові вісті НТУУ "КПІ". — 2008. — No 3. — С. 11—24.
4. Бідюк П.І. Основні етапи побудови і приклади застосування мереж Байєса / Бідюк П.І., Кузнєцова Н.В. // Системні дослідження та інформаційні технології. — 2007. — No 4. — С. 26—39.
5. Бідюк П.І.Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування / Бідюк П.І., Кузнєцова Н. В. // Наукові вісті НТУУ «КПІ». – 2010. – № 1, – С. 42-53.
6. Колбасюк Д.А. Прогнозування темпів приросту ВІЛ/СНІД хворих в Україні на 2016 рік з використанням методів регресійного аналізу / Колбасюк Д.А., Бідюк П.І., Терентьєв О.М., Шумейко О.М. // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта: сб. науч. трудов по материалам междунар. конф., 17-21 мая 2010 г., Евпатория. – Херсон: ХНТУ, 2010. – Т. 2 – С. 80-82.
7. Терентьєв О.М Побудова кредитних скорінгових моделей із використанням аналітичної системи SAS Enterprise Miner / Терентьєв О.М // Системний аналіз та інформаційні технології: матеріали 12-ї Міжнародної науково-технічної конференції SAIT 2010, Київ, 25–29 травня 2010 р. / ННК "ІПСА" НТУУ "КПІ". – К.: ННК "ІПСА" НТУУ "КПІ", 2010. – С. 523.

References

1.Bidyuk P.I. (2009), Assessment individuals models of credit risk, [Modeli otsinky ryzykiv kredytuvannya fizychnykh osib], Cybernetics and computer engineering. - 2007. - No. 153. - pp. 87-95.
2. Terentyev A.N. (2009), Comparison methods of data mining for the estimation individuals creditworthiness, [Sravneniye metodov intellektual'nogo analiza dannykh pri otsenivanii kreditosposobnosti fizicheskikh lits], IKI NASU-NSAU, 2009. - No 5. - pp. 141-149.
3. Kuznetsova N.V. (2008), Systemic approach to credit risk analysis using Bayesian networks [Systemnyy pidkhid do analizu kredytnykh ryzykiv z vykorystannyam merezh Bayyesa], Scientific news NTUU "KPI".- 2008. - No. 3. - pp. 11-24.
4. Bidiuk P.I. (2007), Basic stages of construction and examples using of Bayesian networks, [Osnovni etapy pobudovy i pryklady zastosuvannya merezh Bayyesa], Research systems and information technology. - 2007. - No 4. - pp. 26-39.
5. Bidiuk P.I. (2010), Comparative analysis of characteristics of lending risk assessment models [Porivnyal′nyy analiz kharakterystyk modeley otsinyuvannya ryzykiv kredytuvannya],Scientific news NTUU "KPI".- 2010. - No. 1, - pp. 42-53.

6. Kolbasyuk D.A. (2010), Forecasting growth rate of HIV / AIDS patients in Ukraine in 2016 using regression analysis techniques, [Prohnozuvannya tempiv pryrostu VIL/SNID khvorykh v Ukrayini na 2016 rik z vykorystannyam metodiv rehresiynoho analizu], Intellectual decision-making systems and problems of computing intelligence: Sat. sci. works on the materials of the international. Conf., May 17-21, 2010, Evpatoria. - Kherson: KhNTU, 2010. - T. 2 - pp. 80-82.
7. Terent'ev O. M. (2010), Construction of credit scoring models using the SAS Enterprise Miner analytical system, [Pobudova kredytnykh skorinhovykh modeley iz vykorystannyam analitychnoyi systemy SAS Enterprise Miner], System Analyzes and Information Technologies: Material of the 12th International Scientific and Technical Conference SAIT 2010, Київ, 25-29 травня 2010 р . / NOC "IPSA" NTUU "KPI". - К .: NSC "IPSA" NTUU "KPI", 2010. - P. 523.

Надійшла до редакції 10.10.2017

В.Г. ГУСЬКОВА, П.И. БИДЮК
Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского»

АНАЛИЗ КРЕДИТОСПОСОБНОСТИ ЗАЕМЩИКОВ КРЕДИТОВ С ПОМОЩЬЮ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

В работе предложен подход на основе логистической регрессии для анализа кредитоспособности заемщиков с использованием пакета RStudio. Подход основан на выборе, анализе и применении фактических исходных данных для оценки кредитных рисков в будущем. Использование площади под ROC-кривой и ее значения позволяют найти лучшие полную и сокращенную модели, используемые в ходе работы. Проведен сравнительный анализ полученных моделей и статистических критериев и приведен пример работы с финансовыми данными с использованием данного подхода.

Ключевые слова: логистическая регрессия, анализ кредитоспособности, RStudi, кредитный риск, ROC-кривая, статистические критерии.

V.H. HUSKOVA, P.I. BIDYUK
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

CREDIT ANALYSIS OF BORROWERS USING LOGISTIC REGRESSION

The paper proposes a logistic regression approach for creditworthiness analyzing of borrowers using the RStudio package. The approach is based on the selection, analysis, and application of the actual baseline data to assess credit risks in the future.

Today, the economy, financial and banking sectors are in a rather serious condition, which significantly affects the deterioration of the financial performance of banks, a reduction in the volume of deposits, the provision of loans and a significant decline in the financial stability of banking institutions. As a consequence, this all leads to the fact that a large number of banks become insolvent. This confirms the need to develop relevant measures aimed at overcoming the crisis in the banking sector.

New approaches and methods of risk assessment would allow banks to resume lending and significantly reduce costs. At the stage of collecting statistical data, unknown factors that cause any influence on the result, therefore, financial institutions collect all the information on the basis of which the model is built and the influence of factors on the result is determined. At this stage, the data are analyzed and significant factors are identified.

Using area under the ROC curve and its values allows finding the best complete and shortened models used in the course of work. A comparative analysis of the obtained models and statistical criteria is made and an example of working with financial data using this approach is given

Key words: logistic regression, credit analysis, RStudio, credit risk, ROC-curve, statistical criteria.