МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

АНАЛИЗ МЕТОДОВ ОЦЕНКИ КРЕДИТОСПОСОБНОСТИ КЛИЕНТОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Радченко Екатерины Дмитриевны

Научный руководитель
доцент, к. ф.-м. н. _____ О. А. Мыльцина

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2021

Введение

Оценка кредитоспособности является одним из важнейших процессов принятия банками решений. Этот процесс включает в себя сбор, анализ и классификацию различных данных. Качество банковских кредитов — ключевой фактор, определяющий конкурентоспособность и прибыльность организации. Одним из наиболее важных инструментов моделирования кредитоспособности клиентов является кредитный скоринг.

Кредитный скоринг предполагает использование статистических моделей для определения вероятности того, что потенциальный заёмщик выплатит кредит. Модели кредитного скоринга широко используются для оценки бизнеса, недвижимости и потребительских кредитов.

Кредитный скоринг как метод кредитного анализа используется на протяжении более 60-ти лет. Первая математическая модель для розничного кредита была предложена около 1941 г. в США. Эта модель базировалась на шести параметрах, используемых для регистрации заявок на кредитные карты, таких как должность заявителя и общий стаж на текущем месте работы. Ускорение компьютерных вычислений и рост рынка кредитных карт в 60-х годах способствовали распространению метода.

Большая часть статистических линейных и нелинейных моделей применима для построения рациональной и эффективной системы кредитного скоринга, которую можно использовать с целью вычисления прогнозов. Параметрические техники (WOE (Weight of evidence), корреляционный анализ, регрессионный анализ, дискриминантный анализ, probit-анализ, logit-регрессия, линейное программирование) и непараметрические техники (метод опорных векторов, деревья решений, нейронные сети, метод k-ближайших соседей, генетические алгоритмы и генетическое программирование) — все они так или иначе используются для разработки моделей.

Целью работы является оценка и анализ статистических моделей, пригодных для построения скоринговой системы. Задачи:

1. Изучить ключевые принципы кредитного скоринга;
2. Изучить статистические модели, используемые для построения системы кредитного скоринга;
3. Разработать статистические модели, предсказывающие вероятность де-

фолта по кредиту.

Объект исследования — кредитный скоринг. Предмет исследования — методы, применяемые в кредитном скоринге. Актуальность обусловлена широкой распространенностью скоринговых систем в мировой практике. Результаты анализа статистических моделей могут быть применены в построении реальной скоринговой системы. Оценки моделей могут помочь принять решение в выборе метода моделирования.

Ключевые определяющие кредитного скоринга

Главная задача кредитного скоринга — определить, является ли заемщик «плохим» или «хорошим», или предсказать вероятность того, что заемщик не выплатит кредит. Таким образом, эта задача тесно связана с проблемой классификации.

В качестве объясняющих переменных (предикторов) характеристик при построении моделей используются пол, возраст, семейное положение, количество членов семьи, уровень образования, род деятельности, количество лет, прожитых по текущему адресу, наличие кредитной карты, место работы, величина, продолжительность желаемого займа, наличие недвижимости во владении, наличие автомобиля, ежемесячный доход, цель займа и другие факторы. В некоторых случаях выбор осуществляется при использовании различных техник статистического анализа; например, ступенчатой регрессии или нейронных сетей. Что касается их количества, то оптимального количества не существует, но в некоторых случаях следует учитывать влияние культурных и экономических факторов, присущих региону.

Другим важным моментом является величина выборки. Считается, что чем она больше, тем выше точность скоринговой модели.

Что касается выбора техники, при построении скоринговых моделей используется широкий спектр статистических методов. Любой метод применим для построения эффективной и действенной системы кредитного рейтинга, которую можно использовать в целях прогнозирования. В некоторых случаях выбор техники основан на том, какие объясняющие переменные доступны в наборе.

При моделированиии просрочки и дефолта по потребительским кредитам принято обозначать одну из категорий как хорошую (дефолт по кредиту

не ожидается), другую — как плохую, дефолтную, и обозначать их как 1 и 0 соответственно.

Для решения проблемы пропущенных значений предлагается два подхода — удалить недостающие значения из исходного набора данных или выполнить предварительную обработку для замены отсутствующих значений.

После предварительной подготовки необходимо выбрать процедуру валидации модели.

Оценка моделей может быть осуществлена на основе метрик матрицы ошибок и кривых ROC.

Техники кредитного скоринга

В разделе изучены теоретические основы техник моделирования, используемых для построения моделей.

Линейный регрессионный анализ объясняет связь между переменной отклика и одной или несколькими независимыми переменными и является важнейшим компонентом любого анализа данных.

Методика устанавливает линейную зависимость между характеристиками заемщиков $X = X_1, ..., X_p$ и целевой переменной Y следующим образом:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

где ϵ — случайная ошибка, независимая от X. Предполагается, что ϵ имеет нормальное распределение.

Целью линейного дискриминантного анализа является классификация гетерогенной популяции на однородные подмножества и дальнейший процесс принятия решений по этим подмножествам.

Положим, класс 1 — группа клиентов, которым будет выдан кредит, и класс 2 — клиенты, которым в кредите отказано. При использовании линейного дискриминантного анализа в случае двух классов со средними значениями μ_1 и μ_2 вектор x размерности p будет классифицирован как принадлежащий классу 2, если выполняется условие:

$$x^T(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^T \hat{\sum}^{-1} \hat{\mu}_2 - \hat{\mu}_1^T \hat{\sum}^{-1} \hat{\mu}_1 + \ln(N_1/N) - \ln(N_2/N), \qquad (1)$$

иначе — к классу 1.

Однако оптимальность этого классификационного правила, представленного формулой 1, справедлива лишь в предположении нормально распределенных данных и гомоскедастичности ковариационной матрицы (случается, что данные по кредитам гетероскедастичны). При наличии отклонения от гомоскедастичности требуются квадратичные дискриминантные функции:

$$\delta_k(x) = -\frac{1}{2}\ln|\sum_k| - \frac{1}{2}(x-\mu_k)^T\sum_k^{-1}(x-\mu_k) + \ln\pi_k. \tag{2}$$

Для логистической регрессиии (logit-регрессия) относится к обобщенным линейным моделям и может быть представлена в виде:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j X_j,$$

где $\pi = \mu_Y$ — условное среднее Y, $(\pi/1-\pi)$ — это отношение шансов того, что $Y = 1$, а $\ln(\pi/1-\pi)$ — это логарифм отношения шансов, или логит; ошибки распределяются по логистическому закону. Некоторые исследования признают логистическую регрессию наиболее часто используемой в целях кредитного скоринга.

Модель, ошибки в которой распределены по нормальному закону — probit-регрессия. Эта модель описывается следующим образом:

$$g(\mu_Y) = \Phi^{-1}(p) = \beta_0 + \sum_{i=1}^{m}\beta_i X_i, \tag{3}$$

где $\Phi^{-1}(p)$ используется в качестве связующей функции — определяет, probit-трансформацию зависимой переменной.

Обобщенная аддитивная модель принимает вид:

$$g(\mu_Y) = \beta_0 + \sum_{i=0}^{p} f_i(X_i), \tag{4}$$

где $X_1, X_2, \ldots, X_p$ — предикторы, Y — переменная отклика и f_i — непара-

метрическая гладкая функция от для предиктора X_i.

Метод к-ближайших соседей берет за основу следующее правило: объект считается принадлежащим к какому-либо классу, если к нему принадлежит большинство его соседей. При использовании этого метода решающее значение имеет выбор метрики. Обычно метрикой является стандартная евкдидова норма, представленная формулой:

$$d_1(x, y) = (x - y)^T (x - y)^{1/2}, \tag{5}$$

где x и y — точки в пространстве.

Наиболее распространенной в кредитном скоринге, а также в прочих задачах, касающихся классификации, архитектурой нейронных сетей является многослойный персептрон. Сеть состоит из входного слоя, одного или нескольких скрытых слоев и выходного слоя; каждый слой состоит из нескольких нейронов. Входной слой соединен со скрытым слоем, который соединен, в свою очередь, с выходным слоем.

Метод опорных векторов можно представить как поверхность, которая образует границу между точками данных, нанесенными на график в многомерном пространстве, описывающем примеры и значения их признаков. Цель SVM состоит в том, чтобы построить плоскую границу — гиперплоскость, которая бы делила пространство таким образом, чтобы по обеим ее сторонам образовались однородные группы.

Алгоритм CART (classification and regression trees) — методы классификации и регрессии с использованием дерева решений. Это методика обучения, основанная на деревьях решений, которая возвращает классификационные или регрессионные деревья.

Моделирование

Язык программирования, выбранный для осуществления задачи моделирования — R.

Для построения моделей в настоящей работе задействовано множество пакетов, зарекомендовавших себя в анализе данных. Среди тех, что используются на каждом из этапов моделирования — `tidyverse` и `caret`.

Для построения моделей использованы данные, содержащие 12 коло-

нок. В качестве зависимой переменной Y будет использоваться характеристика SeriousDlqin2yrs, принимающая значение 1 в случае, если у плательщика ожидается дефолт по кредиту, и 0, если кредит будет успешно выплачен. Перед моделированием данные очищены от выбросов, распределение данных приведено к нормальному для использования в моделях, на чью производительность оно может повлиять.

Оценка обучающей способности алгоритмов, используемых в работе, будет осуществляться при помощи техники многократной k-блочной кросс-валидации с параметром $k = 10$ и 10 повторениями.

Функции для построения моделей содержатся в пакетах `MASS`, `glm`, `mgcv`, `class`, `rpart`, `nnet`, `e1071`. Построены следующие модели:

— Линейная дискриминантная и квадратичная дискриминантная модели;
— Logit и probit модели;
— Обобщенная аддитивная модель;
— Модель на основе метода k-ближайших соседей;
— Модель на основе многослойного персептрона;
— Модель на основе метода опорных векторов;
— Модель на основе алгоритма CART.

Анализ моделей

Анализ моделей осуществляется в соответствии с матрицей ошибок и кривыми ROC. Показатели точности, чувствительности, специфичности и ошибки классификации моделей представлены в таблице 1.

Согласно таблице 1, модели дискриминантного анализа — линейная (LDA) и квадратичная (QDA) — обладают достаточно высокой точностью. При этом модель, построенная на основе линейного дискриминантного анализа, классифицирует наблюдения, в которых кредит будет выплачен, с корректностью свыше 99%, однако имеет низкий показатель специфичности. Наиболее верно классифицирует наблюдения, в которых кредит не будет выплачен, модель, основанная на методе квадратичного дискриминантного анализа — этот показатель наибольший среди всех построенных моделей. При использовании QDA лишь пятая часть клиентов, невыполнивших обязательства по кредиту, классифицирована ошибочно. Тем не менее, чувствительность и точность LDA превышают аналогичные показатели QDA, и показа-

Таблица 1 – Показатели моделей

	Точность	Чувствительность	Специфичность	Ошибка классификации
LDA	92,13%	99,45%	65,77%	7,87%
QDA	90,81%	93,75%	80,24%	9,19%
Logit	93,32%	98,76%	73,73%	6,68%
Probit	93,18%	98,82%	72,86%	6,5%
GAM	93,87%	99,95%	71,98%	6,13%
KNN	92,35%	98,37%	71,29%	7,65%
NN	87,12%	98,91%	44,67%	12,88%
SVM	93,38%	98,89%	73,54%	6,62%
CART	83,55%	93,97%	47,39%	16,5%

тель чувствительности QDA является наиболее низким в сравнении с прочими моделями.

Показатели обобщенных линейных моделей близки, разница в точности и ошибки классификации между логистической (logit) и probit моделями менее одного процента. С учётом всех рассмотренных методов, обобщенные линейные модели имеют достаточно высокие величины метрик точности, чувствительности, специфичности и невысокую ошибку классификации.

Обобщенная аддитивная модель (GAM) имеет наиболее высокую точность и обладает наибольшей чувствительностью.

Метод k-ближайших соседей (KNN) по значению специфичности близок к обобщенной аддитивной модели, среди непараметрических методов (нейронные сети (NN), метод опорных векторов (SVM), алгоритм CART) обладает наибольшей точностью.

Нейронная сеть неверно классифицирует более 10% наблюдений, при этом обладает чувствительностью, немногим уступающей обобщенной аддитивной модели. Однако специфичность модели низкая: более половины невыплаченных кредитов были классифицированы неверно.

Метод опорных векторов даёт показатели, сравнимые с теми, что получены при помощи обобщенных линейных моделей.

Результаты работы алгоритма CARD среди всех рассмотренных моделей имеют наиболее низкую точность. Модель обладает низкой специфично-

стью, однако она несколько выше, чем специфичность предыдущей модели.

Важно отметить, что все непараметрические техники, как ожидалось, требуют достаточно большой вычислительной мощности, особенно при учёте большого количества данных, что делает процесс обучения моделей времязатратным.

Визуализировать производительность моделей можно с помощью кривых ROC. На рисунке 1 кривые ROC для каждой из моделей размещены на одном графике.

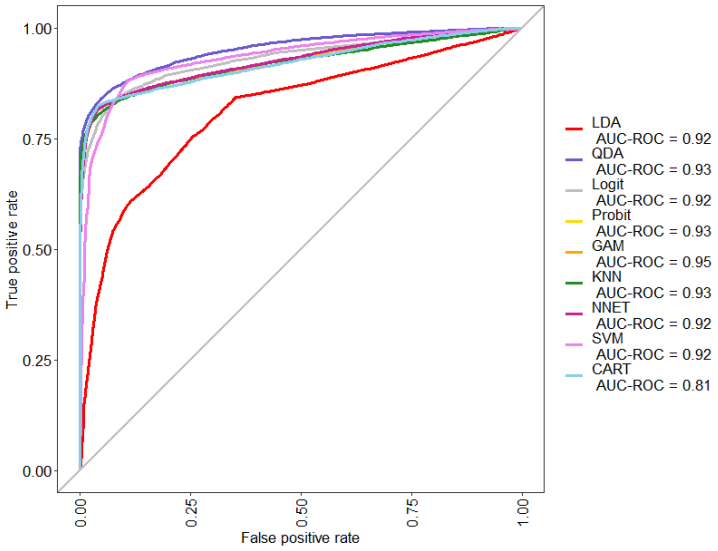

Рисунок 1 – Кривые ROC для рассматриваемых моделей

Следуя рисунку 1, наиболее удаленной от главной диагонали и приближенной к левому верхнему углу кривой, а также наимеющей наибольшую величину AUC является кривая, относящаяся к квадратичной дискриминантной модели. Моделью достигается наибольшее соотношение чувствительности и специфичности; метод опорных векторов также достигает этой точки. В целом кривые моделей достаточно близки, что говорит об их применимости для решения задачи классификации кредитов с примерно одинаковой производительностью.

Таким образом, выбор модели в основном зависит о целей, преследуемых при построении. В случае, когда важно допустить как можно меньше ошибок в выдаче кредитов неблагонадежным заёмщикам, результативна квадратичная дискриминантная модель. С учётом этого параметра неудачны

модели на основе нейронных сетей и алгоритма CART. Обобщенная аддитивная модель актуальна в случае, когда существует заинтересованность в выдаче большего количества кредитов, лицам, которые способны их выплатить. В целом все приведенные модели имеют хорошие показатели точности (более 80%) и чувствительности (более 90%), 6 из 9 моделей обладают специфичностью выше 70%. При этом непараметрические методы, несмотря на простоту реализации, требуют большого временного ресурса, что создаёт неудобство в обучении моделей на большом количестве данных.

Заключение

Целью работы бакалаврской работы являлись оценка и анализ статистических моделей, пригодных для построения скоринговой системы. Объект исследования — кредитный скоринг. Предмет исследования — методы, применяемые в кредитном скоринге.

В ходе работы изучены теоретические основы кредитного скоринга, этапы построения скоринговых систем, различные статистические методы, популярные в области кредитного скоринга. Среди изученных техник моделирования — параметрические и непараметрические методы среди которых дискриминантный анализ, обобщенные линейные модели, обобщенные аддитивные модели, метод k-ближайших соседей, нейронные сети, метод опорных векторов и алгоритм CART. На языке программирования R, базируясь на открытых данных, осуществлен процесс моделирования изученных методов. В процессе написания программного кода задействованно множество библиотек, служащих в работе над обучением и тестированием моделей, а также их построением.

При анализе полученных моделей установлено:

— Наибольшей точностью обладает обобщенная аддитивная модель;
— Наиболее верно классифицирует наблюдения, претерпевшие дефолт, квадритичная дискриминантная модель;
— Наиболее верно классифицирует наблюдения с успешной выплатой кредитов обобщенная аддитивная модель;
— Наибольшая ошибка классификации — у модели, построенной при помощи алгоритма CARD.