**Санкт-Петербургский государственный университет**



Факультет прикладной математики-процессов управления

Магистерская диссертация

**Построение скоринговых карт с использованием модели логистической регрессии**

Направление 03.04.01 «Прикладные математика и физика»

ООП ВМ.5521.2018 «Математические и информационные технологии»

Выполнил: Н.Н. Поздняков

Научный руководитель:\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_Л.В.Владимирова

Члены комиссии: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_О.И.Дривотин

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Н.С. Едаменко

Санкт-Петербург

2021

Оглавление

[Введение 3](#_Toc71452827)

[Глава I. Логит-модель 7](#_Toc71452828)

[Глава II. Статистическая обработка полученных результатов в логит-модели 16](#_Toc71452829)

[Глава III. ROC-анализ 21](#_Toc71452830)

[Глава IV. Численные результаты 27](#_Toc71452831)

[Заключение. 46](#_Toc71452832)

[Список литературы 47](#_Toc71452833)

# Введение

В настоящее время скоринг становится все более популярным при различных видах кредита, а также и в других областях.

В банковской системе кредитный скоринг можно опередить как метод начисления потенциальным заемщикам определенного количества баллов на основе информации о его социально-демографическом положении, кредитной истории, параметрах запрашиваемого кредита, и принятие решения о выдаче или об отказе в кредите на основе набранного суммарного количества баллов.

Если обратиться к истории, то скоринг, по существу, является методом классификации всей интересующей нас популяции на различные группы, когда нам неизвестна характеристика, которая разделяет эти группы (вернет клиент кредит или нет), но зато известны другие характеристики, связанные с интересующей нас. В статистике идеи классификации популяции на группы были разработаны Фишером в 1936 г. на примере растений. В 1941 г. Дэвид Дюран впервые применил данную методику к классификации кредитов на «плохие» и «хорошие». По времени это совпало со Второй мировой войной, когда почти все кредитные аналитики были призваны на фронт, и банки столкнулись с необходимостью срочной замены этих специалистов. Банки заставили своих аналитиков перед уходом написать свод правил, которыми следовало руководствоваться при принятии решения о выдаче кредита, чтобы анализ мог проводиться неспециалистами. Это и был как бы прообраз будущих экспертных систем.

В начале 50-х гг. в Сан-Франциско образовалась первая консалтинговая фирма в области скоринга – Fair Issac, которая до сих пор является лидером среди разработчиков скоринговых систем.

Одна из проблем заключается в том, что люди с течением времени меняются, меняются и социально-экономические условия, влияющие на поведение людей. Поэтому скоринговые модели необходимо разрабатывать на выборке из наиболее «свежих» клиентов, периодически проверять качество работы системы и, когда качество ухудшается, разрабатывать новую модель. На Западе новая модель разрабатывается в среднем раз в полтора года, период между заменой модели может варьироваться в зависимости от того, насколько стабильной была экономика в это время.

Для России, вероятно, максимальным периодом будет полгода, да и то при условии, что в этот период не произойдет никаких кардинальных потрясений.

В настоящее время ведутся исследования того, как вводить социально-экономические характеристики в модель с тем, чтобы она служила дольше.

Широкое применение скоринга началось с распространением кредитных карточек. При том количестве людей, которые ежедневно обращались за кредитными карточками, банкам ничего другого не оставалось, как автоматизировать процесс принятия решений по выдаче кредита. Однако очень скоро они оценили не только быстроту обработки заявлений на выдачу кредита, но и качество оценки риска. По данным некоторых исследований, после внедрения скоринг-систем уровень безнадежного долга сокращался до 50% [10, 14]. .

В 1974 г. в США был принят Закон о предоставлении равных возможностей на получение кредита, который запрещал отказывать в выдаче кредита на основании следующих характеристик: раса, цвет кожи, национальное происхождение, возраст, пол, семейное положение, религия, получение социальных пособий, отстаивание прав потребителей.

В Великобритании законодательство допускает использование информации о возрасте и семейном положении, но зато запрещает принимать во внимание какие-либо физические увечья и недостатки (инвалидность).

Для кредитных организаций использование скоринговых систем стало доказательством исполнения этих антидискриминационных законов – у компьютера нет предубеждений.

Помимо установления принципов равноправия в области кредитования, кредитное законодательство США, как и Закон о потребительском кредите, принятый в Великобритании в том же 1974 г., имели важное значение для формирования службы кредитных бюро. В таких бюро записывается кредитная история всех людей, когда-либо обращавшихся за ссудой в любую кредитную организацию страны.

В кредитных бюро содержатся следующие виды данных:

* социально-демографические характеристики;
* судебные решения (в случае передачи дел о востребовании задолженности по кредиту в суд);
* информация о банкротствах;
* данные об индивидуальных заемщиках, получаемые от кредитных организаций по принципу «ты - мне, я - тебе», т. е. банк может получать информацию о клиентах других банков, только если сам поставляет аналогичную информацию.

Объем и характер информации, хранящейся в бюро, строго регулируется законодательством каждой страны. В «Банковских технологиях» уже была публикация о кредитных бюро в сентябре 1999 г. – «Вопросы учреждения кредитного бюро в России».

Хотелось бы добавить, что существуют транснациональные коммерческие компании, такие как Experian, Equifax, TransUnion, Scorex. Эти компании сами используют скоринговые системы, и во многих случаях продают клиентам не «сырую» информацию, а уже готовый интегральный показатель, который вводится в автоматизированную систему кредитной организации.

Значение кредитных бюро чрезвычайно велико, их существование позволяет кредитным организациям выдавать ссуды клиентам, которые ранее в этой организации не обслуживались. Кроме того, общепризнанной является ценность предыдущей кредитной истории для прогнозирования вероятности дефолта.

Еще чуть более десятка лет назад обычный человек мог и не мечтать о том, чтобы получить кредит в банке без имущественного залога или поручительства третьего лица. А иногда банковские структуры требовали то и другое вместе. В этой ситуации банкиры чувствовали себя полностью защищенными от каких-либо неприятностей и наглели до безобразия. Всем памятны кредитные ставки в 50% и более в сочетании с жесткими санкциями за малейшую промашку, которую мог бы совершить клиент.

Время и рыночные отношения внесли свои коррективы. Ситуация в корне изменилась. Сегодня банкам приходится бороться за клиента, они стараются заманить потребителя привлекательными ставками и условиями. Речь уже не идет о необходимости предоставления залога или поручителя. Зачастую достаточно внутри гражданского паспорта и второго документа. Таким образом, кредитное обеспечение невероятно снизилось. Одним из важнейших показателей, привлекательного для клиента, стало быстрое принятие решения о выдаче кредита.

Если говорить о преимуществах, будущем и актуальности, то скоринг, как метод оценки кредитного риска, позволяет максимально ускорить процесс независимого принятия кредитного решения за счет практически полной автоматизации процесса расчета и максимальной минимизации влияния человеческого фактора. Полностью сделать процесс независимым, невозможно по причине того, что исходные данные вводятся оператором, конечное заключение готовится кредитным аналитиком, а принимается решение, уполномоченным на то менеджером.

Требования рынка корректируют все объективные или субъективные недоработки и недостатки, что, в конечном итоге приведет к повсеместному применению данной, передовой системы.

В настоящей работе проводится **кредитный скоринг** (от англ. score «оценка»), то есть система оценки кредитоспособности лица, на основе численных статистических методов.

Для этого была построена оптимальная логит-модель, которая определяла вероятность выдачи кредита на основе обучающей выборки.

Эта сложная задача решалась в 3 этапа.

Первой простейшей задачей являлась задача с одним предиктором. Она наглядно показала суть задачи. Однако такая далека от жизненной ситуации, когда можно принимать решение только на основании величины заработной платы.

Поэтому далее перешли ко второй задаче с 6 предикторами.

Благодаря проверке гипотез о значимости коэффициентов, было получено оптимальное число параметров и решена третья задача с этим числом параметров.

# Глава I. Логит-модель

Суть логистической регрессии заключается в анализе связи между несколькими независимыми переменными, называемыми регрессорами или предикторами, и зависимой переменной – откликом, причем зависимая переменная является бинарной, то есть *принимает только два значения*.

* 1. **Логистическая регрессия**

В данной работе рассматривается задача о построении логистической регрессии с бинарной зависимой переменной. Логит-преобразование позволяет получать значение величины, принимающей значение в (0,1), и интерпретируется как функция распределения некоторой случайной величины.

Значение прогнозируемого отклика, близкое к 0, говорит о малой вероятности наступления интересующего нас исхода, а значение, близкое к 1, говорит в пользу высокой вероятности наступления этого события.

Пусть величины являются предикторами (характеристиками) клиентов. Множественная регрессия

, (1.1)

где вектор неизвестных параметров, приводит к модели с предсказываемыми значениями большими 1 и меньшими 0. Но такие значения вообще не допустимы для первоначальной задачи.

Для решения проблемы задача регрессии может быть сформулирована иначе: вместо предсказания бинарной переменной , будем предсказывать непрерывную переменную со значениями на отрезке при любых значениях независимых переменных. Это достигается применением следующей регрессионной логит-функции



Рис.1.1. Логит-функция

. (1.2)

Она изображена на рис. 1.1. В силу (1.1), функция (1.2) зависит от вектора предикторов

и от вектора неизвестных параметров . Будем обозначать ее следующим образом

.

Следовательно, в основе логистической регрессионной модели прогнозируется *не само значение переменной отклика , а вероятность того, что переменная отклика принимает значение 1.*

**Цель работы**

***Требуется подобрать неизвестные параметры логистической регрессионной модели таким образом, чтобы по логит-модели можно было бы прогнозировать решение о выдаче или отказе кредита данному клиенту.***

Сначала для решения задачи подберем подходящий критерий. Таким критерием является функция правдоподобия (likelihood function), так как она содержит всю имеющуюся информацию, используемую при решении поставленной задачи.

* 1. **Функция правдоподобия**

Пусть величины являются предикторами (характеристиками) клиентов и имеется наблюдений – зависимая переменная, которая может принимать значения ноль и единица, .

Метод максимального правдоподобия (***m****aximum* ***l****ikelihood* ***e****stimation*) – это метод оценивания неизвестных параметра путём максимизации [функции правдоподобия](https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F_%D0%BF%D1%80%D0%B0%D0%B2%D0%B4%D0%BE%D0%BF%D0%BE%D0%B4%D0%BE%D0%B1%D0%B8%D1%8F) [1].

Будем полагать, что вероятность наступления события  равна , а вероятность наступления события  равна . Таким образом, при определении вероятности наступления события , мы имеем распределение Бернулли [4] и функция распределения  будет иметь вид . Запишем этот факт для каждого наблюдения

 (1.3)

и составим функцию правдоподобия

 . (1.4)

Заметим, что

,

,

где .

Тогда функцию правдоподобия (1.4) можно упростить

. (1.5)

Для нахождения оптимальных параметров будем оптимизировать функцию (1.5), имея данные об клиентах: .

* Для упрощения вычислений можно рассмотреть , так как оптимальные значения от функции и ее логарифма совпадают.

, (1.6)

где .

* 1. **Методы оптимизации для получения оптимальных коэффициентов**

Целевая функция (1.6) зависит от -ой переменной . Для нахождения максимума будем использовать комбинацию двух методов –случайный поиск и градиентный. Такая комбинация является эффективной. Доказано [3,5], что случайный поиск находит глобальный экстремум с точностью , где – число пробных точек, и служит “разведочным аппаратом”. Градиентный метод уточняет полученный результат.

***А).******Случайный поиск***

Пусть дана функция  определенная в некоторой ограниченной области  m+1-мерного евклидова пространства – ; . Рассмотрим следующую экстремальную задачу: требуется найти точку , удовлетворяющую соотношению

. (1.7)

Точку , удовлетворяющую (1.7), будем называть точкой глобального экстремума (или экстремальной точкой) функции  в области .

Простейший случайный поиск основан на «просмотре» многомерной области  с помощью равномерно распределённых в этой области случайных точек [3,5].

**Алгоритм простейшего случайного поиска**

Построим последовательность

, (1.8)

независимых реализаций равномерно распределенных в области  случайных векторов.

**Замечание.** Если область  является -мерным параллелепипедом

, (1.9)

где некоторые заданные числа. Тогда [3]

. (1.10)

В (1.10) величины являются независимыми равномерно распределенными в промежутке .

Положим

,

, , , (1.11)

. (1.12)

*В результате процедуры получаем вектор  из множества пробных точек (1.8), для которого значение целевой функции  наибольшее.*

Покажем, что при неограниченном увеличении числа испытаний вероятность попадания хотя бы одной пробной точки в окрестность экстремальной точки множество стремится к единице [3,5]. Действительно, пусть . Вероятность попадания пробной точки при одном испытании в множество  равна: . Каждая точка  последовательности (1.8) с вероятностью  попадает в множество  и с вероятностью  не попадает туда (распределение Бернулли). Найдем вероятность  того, что хотя бы одна из пробных точек (1.8) попала в множество .

. (1.13)

Очевидно, что .

Таким образом, пробная точка, попавшая в множество , является приближением точки супремума .

Итак, доказана следующая теорема.

**Теорема 1.1** [5] *Пусть функция  ограничена сверху и кусочно-непрерывна в области * (). *Вероятность того, что хотя бы одна точка последовательности (1.8) при  испытаниях попадёт в окрестность точки  супремума* *функции , стремится к 1 при .*

Утверждение теоремы означает, что при сделанных предположениях имеет место ***сходимость простейшего случайного поиска*** глобального экстремума функции *.*

Соотношение (1.13) позволяет оценить число испытаний, необходимое для достижения требуемой точности отыскания точки глобального экстремума . Пусть задано достаточно малое число . Запишем условие того, что вероятность попадания хотя бы одной пробной точки при  испытаниях в множество  не меньше , где :

 .

Тогда . Следовательно,

.

Минимальное число испытаний, удовлетворяющее последнему условию, равно

+1, (1.14)

где  – целая часть . Итак, при  вероятность попадания хотя бы одной пробной точки (3.2) в множество  положительной меры  не меньше .

***А). Градиентный метод***

Будем рассматривать оптимизационную задачу (1.7), то есть искать экстремальную точку в области .

Для приближенного поиска точки экстремума с использованием направления поиска по градиенту часто используют последовательность, построенную специальным способом и удовлетворяющей некоторым условиям.

**Определение**. Будем говорить, что последовательность  является *максимизирующей последовательностью* для  в области , если выполняется следующее равенство

. (1.15)

**Определение.** Процесс построения максимизирующей последовательности  будем называть *релаксационным*, если

. (1.16)

Выбор способа построения  существенно зависит от свойств , от информации, которая используется на каждой итерации, а также от технических средств, которыми располагаем для осуществления вычислительных процедур.

Методы можно подразделить на три группы [7]:

1. методы нулевого порядка, которые используют только значение ;
2. методы первого порядка, использующие и значения  и ее первую производную ;
3. методы второго порядка, в котором используются значения , , .

**Общая схема направленных методов поиска экстремума функции**

Методы поиска экстремума функции различаются либо выбором направления, либо способом движения вдоль направления спуска. В качестве начального приближения выбирается любая точка . Последовательные приближения строятся по следующей схеме:

1. в точке  выбирается направление возрастания функции
2. приближение находится по формуле

(1.17)

Здесь величина  *с точностью до множителя * определяет расстояние между точками ,  и называется *длиной шага* или просто *шагом*.

Вектор *направления спуска* выбирается таким образом, чтобы для малых значений  выполнялось неравенство

(1.18)

***Замечание 1.*** Градиентный спуск проводится до тех пор, пока

(1.19)

или

. (1.20)

Здесь – заданные маленькие величины.

В (1.19) норма вектора находится по формуле

.

***Замечание 2.*** Часто полагают величину шага не зависящей от номера итерации , то есть считают равной некоторой постоянной величине.

***Замечание 3.*** В учебнике [7] можно найти ряд других релаксационных методов поиска экстремума функций, например, метод Ньютона, покоординатный спуск и др..

***С. Получение вектора градиентов целевой функции***

Целевая функция имеет вид (1.6)

т.к.

.

Тогда вектор градиента логарифма функции правдоподобия (целевая функция) вычисляется следующим образом:

. (1.21)

Начальная точка градиентного метода

выбирается равной приближенному экстремуму, полученному простейшим случайным поиском.

Шаг может быть, как постоянным, так и разным на каждой итерации. Рассмотрим переменный шаг Его можно находить, используя один из методов одномерной максимизации

, (1.22)

либо на каждой итерации делим шаг пополам, если не произошло увеличение целевой функции с шагом .

В случае (11) градиентный метод называется *методом наискорейшего спуска*.

Параметры метода:

шаг на каждой итерации,

– начальная точка,

– маленькая величина для определения останова алгоритма:

Если

, (1.23)

то можно заканчивать поиск.

Процесс градиентного поиска продолжается до получения наибольшего значения целевой функции.

* Будем полагать, что момент окончания поиска наступит тогда, когда движение из полученной точки с любым шагом не приводит к возрастанию значения целевой функции.
* Если минимум функции достигается внутри рассматриваемой области, то в этой точке градиент равен нулю, что также может служить сигналом об окончании процесса оптимизации.
* Также, момент окончания поиска можно определить проверкой близости значений целевой функции на двух последовательных итерациях.

# Глава II. Статистическая обработка полученных результатов в логит-модели

**2.1. Вероятностная модель**

При решении регрессионной задачи на основе выбранных характеристик клиентов делается вывод о выдачи кредита или отказе. Однако в жизни возникают также и неучтенные факторы, совокупное влияние которых моделируется случайной компонентой. Это возможно интерпретировать ситуацию следующим образом. Будем предполагать, что существует некоторая переменная , связанная с независимыми переменными-предикторами линейным регрессионным уравнением

, (2.1)

где случайные компоненты взаимно независимы и нормально распределены с нулевым математическим ожиданием и дисперсией . Переменная является ненаблюдаемой, а решение, соответствующее значению , принимается тогда, когда превосходит некоторое пороговое значение. С учетом соотношения (2.1) оценки параметров логит-модели, найденные методом максимального правдоподобия, асимптотически подчиняются многомерному нормальному распределению [1,6,11]

, (2.2)

где – ковариационная матрица вектора и равна

. (2.3)

В равенстве (2.3) величина

где

. (2.4)

*является несмещенной оценкой* одиночной дисперсии .

Предполагаем , вхдящие в (2.1) нормально распределенными, независимыми, центрированными и некоррелированными:

, ,

то есть . Тогда можно использовать формулу моделирования [1,3,5]

– независимые равномерно распределенные числа в (0,1), .

Матрица размерности ( предикторов для клиентов)

и

. (2.5)

Это симметричная матрица относительно диагонали размерности

. Далее нужно найти обратную матрицу и использовать ее диагональные элементы.

**2.2. Построение доверительных интервалов**

Нахождение доверительного интервала для оценки параметра играет большую роль. Зная доверительный интервал, можно сказать, что *с заданной вероятностью найденный интервал покрывает неизвестный параметр*.

**Доверительным интервалом** [1,2,4,6]с уровнем доверия неизвестного параметра распределения называют интервал

,

обладающий следующим свойством

. (2.7)

Соотношение (2.7) можно еще представить следующим образом

или

. (2.8)

Соотношение (2.7) можно понимать следующим образом: вероятность того, что интервал

(2.9)

заключает в себе неизвестный параметр , равна . Таким образом, ***доверительный интервал (2.9) покрывает параметр с вероятностью .***

Рассмотрим нормированную случайную величину (статистику)

, (2.10)

где величина в знаменателе ***является оценкой неизвестной дисперсии для , поэтому статистика (2.10) подчиняется распределению***

***Стьюдента*** с степенями свободы [1,2,4]. Тогда

. (2.11)

В (2.11) величина находится по таблице распределения Стьюдента по значениям и . Функция под интегралом в (2.11) является плотность распределения вероятностей Стьюдента. Раскрывая неравенство под вероятностью в (2.11), находим доверительный интервал для :

. (2.12)

– квантиль распределения Стьюдента с степенями свободы уровня доверия .

**2.3. Проверка статистических гипотез о значимости параметров логит-модели**

Для логит-модели проверка гипотезы о значимости одного параметра может проводиться с помощью следующей статистики

, (2.13)

если требуется проверить . *Эта проверка очень важна, так как дает возможность уменьшить число параметров*.

Как было отмечено ранее, оценки параметров логит-модели , найденные методом максимального правдоподобия, асимптотически подчиняются многомерному нормальному распределению (2.2).

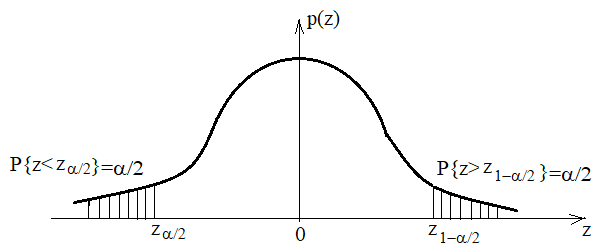


Рис.2.1. Критическая область (заштрихованная область)

**Алгоритм проверки статистической гипотезы**

1. Выдвигаем нулевую гипотезу . Альтернативная гипотеза – это .
2. Зададим уровень значимости .
3. Находим критическую область

(2.14)

где – квантиль стандартного нормального распределения уровня .

1. Вычислим значение .
2. Если значение статистики попадает в критическую область, то есть событие маловероятно, тогда гипотеза отвергается. Иначе нет снований ее отвергнуть при уровне значимости .

*Применяя этот алгоритм при , можно выявить значимые коэффициенты и по ним, соответственно, определить значимые предикторы.*

# Глава III. ROC-анализ

ROC-анализ (Receiver Operator Characteristic) это особый метод статистического анализа, пригодный осуществления бинарной классификации, т.е. выявляющих наличие некоего состояния (условно обозначаемое «1») либо его отсутствие (условно обозначаемое «0»). Соответственно, два варианта выявляемого состояния должны быть *взаимоисключающими*– либо клиент получает кредит либо нет. Таким образом, ROC-анализ оперирует двумя классами событий – с *положительными исходами* и с *отрицательными исходами*.

* 1. ***Четырехпольная таблица сопряженности (confusion matrix)***

ROC-кривая показывает зависимость количества верно классифицированных положительных исходов от количества неверно классифицированных отрицательных исходов [8,9,11-13]. В терминологии ROC-анализа первые называются *истинно положительным множеством*, вторые – *ложно отрицательным множеством*. При этом предполагается, что у классификатора (исследуемой независимой переменной) имеется некоторый параметр, варьируя который, мы будем получать то или иное разбиение значений вероятностей на два вышеупомянутых класса. Этот параметр называют порогом, или точкой отсечения (cut-off value). В зависимости от него будут получаться различные величины *ошибок I и II рода*.

Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу сопряженности (confusion matrix), называемой матрицей классов или четырехпольной таблицей (табл. 3.1), которая строится на основе результатов классификации с использованием регрессионной модели и принадлежностью наблюдений к двум вышеописанным классам.

|  |  |  |
| --- | --- | --- |
| Фактическое состояние объектов | Решение по тестируемому методу | |
| Положительное | Отрицательное |
| Положительное |  |  |
| Отрицательное |  |  |

*Таблица 3.1.* Матрица классов

В бинарной классификации каждое предсказание может иметь четыре исхода:

TP (*True Positives*) – верно классифицированные положительные примеры (истинно положительных случаи),

TN – (*True Negatives*) – верно классифицированные отрицательные примеры (истинно отрицательные случаи),

FP – (*False Positives*) – отрицательные примеры, классифицированные как положительные (ошибка II рода). Это ложное обнаружение, так как при отсутствии события ошибочно выносится решение о его присутствии (ложноположительных случаи),

FN – (*False Negatives*) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемы “ложный пропуск” – это когда интересующее нас событие ошибочно не обнаруживается (ложноотрицательные примеры).

* 1. ***Специфичность, чувствительность и точность логит-модели***

С помощью матрицы классов можно оценить точность модели, которая вычисляется по формуле:

. (3.1)

ROC-анализ представляет собой графическую методику оценивания эффективности модели с помощью двух показателей – специфичности и чувствительности.

**Чувствительность** (sensitivity) является отношением числа истинно положительных наблюдений к числу фактически положительных наблюдений:

. (3.2)

**Специфичность** (specificity) определяется как отношение числа истинно отрицательных наблюдений к числу фактически отрицательных наблюдений:

. (3.3)

В логистической регрессионной модели прогноз положительного события дается при его вероятности  и отрицательного – при его вероятности . Таким образом, порог по разделению двух классов (cutoff) равен .

Величина cutoff может меняться в пределах от 0 до 1. Построим sensitivity и specificity в зависимости от cutoff по формулам (3.2) и (3.3) соответственно. На рис. 3.1 кривая красного цвета отражает изменение значения sensitivity в зависимости от cutoff, а кривая зеленого цвета – значения specificity.

При этом следует учитывать, что при повышении чувствительности неизбежно падает специфичность, и наоборот – это вытекает из самой сущности ROC-анализа. Уменьшение величины порога отсечения повышает чувствительность, увеличение – специфичность.

* 1. ***Выбор порога отсечения по разделению двух классов***

Порог отсечения нужен для того, чтобы применять диагностический тест на практике – относить новые случаи к одному из двух классов. Для определения оптимального порога отсечения нужно задать критерий его определения.



Рис. 3.1. Изменение чувствительности и специфичности от порога отсечения

Следует заметить, что при возрастании специфичности чувствительность убывает, поскольку **при возрастании порога отсечения количество истинно классифицированных отрицательных случаев растет, а количество истинно классифицированных положительных случаев снижается.**

Существуют разные критерии выбора порога отсечения, связанные с оптимальной стратегией конкретной задачи. В качестве критериев могут выступать следующие критерии:

* Требование минимальной величины чувствительности (специфичности) модели
* Требование максимальной суммарной чувствительности и специфичности модели
* Требование баланса между чувствительностью и специфичностью
* Требование минимизации цены ошибок (когда ошибкам I и II рода назначается вес)

**Будем рассматривать в качестве порога отсечения точку пересечения кривых чувствительности и специфичности.** Эта точка дает значение оптимального значения отсечения, то есть оптимального порога разделения двух классов, и уравновешивает показатели чувствительности, специфичности, качества модели.

***3.4. ROC-кривая и показатель качества логистической модели***

Построим ROC-кривую, которая отражает качество логистической модели. Диаграмма строится в координатах (1-Specifity) и Sensitivity. ROC-кривая является параметрической кривой. Специфичность и чувствительность это независимые характеристики и зависят от одного параметра – порога отсечения. Порог отсечения, как и вероятность, варьируется от 0 до 1 с шагом (например. 0.01). Меняя порог отсечения, получаем набор точек, которые и изображены на Рис. 3.2.

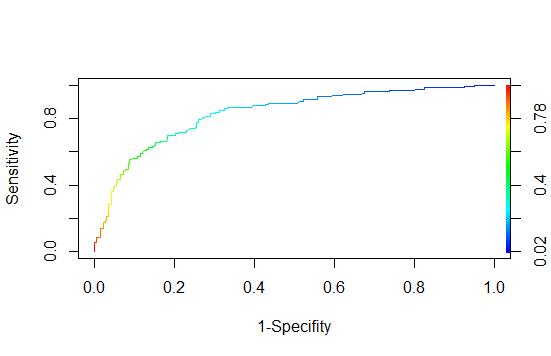


Рис. 3.2. ROC-кривая

На рис. 3.2 цвет графика соответствует значению параметра (порога отсечения). Оптимальный параметр должен максимизировать чувствительность и специфичность. Следовательно, оптимальный параметр соответствует точке на графике, которая максимально приближена к верхнему левому углу пространства построения.

Также для анализа модели интересен показатель AUC (area under ROC curve). Этот показатель определяется площадью под кривой (может быть определен методом трапеций). Максимальное значение AUC может быть равным 1, когда 1-Specifity и Sensitivity будут равны 1. В таблице 3.2 приведены показатели AUC с соответствующими оценками модели[8].

|  |  |
| --- | --- |
| AUC | Качество модели |
| 0.9-1.0 | Отличное |
| 0.8-0.9 | Очень хорошее |
| 0.7-0.8 | Хорошее |
| 0.6-0.7 | Среднее |
| 0.5-0.6 | Плохое |

*Таблица 3.2. Показатель AUC*

Следует отметить, что показатель AUC применяют для сравнения нескольких моделей, он не несет никакой информации о чувствительности и специфичности.

# Глава IV. Численные результаты

**Задача 1.** Для наглядности рассмотрим задачу построения скоринга с одним предиктором ()

В этом случае функция регрессии с одним предиктором имеет вид

, (4.1)

где – число клиентов, а – заработная плата каждого (один предиктор).

**Простейший случайный поиск**

В этом случае целевая функция зависит от и имеет вид

(4.2)

Рассматриваются данные 50 клиентов (обучающая выборка)

Табл. 4.1 Обучающая выборка с одним предиктором

|  |  |  |
| --- | --- | --- |
| **№** | **ЗП, т.р.** | **Отклик** |
| 1 | 55 | 0 |
| 2 | 40 | 0 |
| 3 | 75 | 1 |
| 4 | 55 | 1 |
| 5 | 45 | 1 |
| 6 | 40 | 0 |
| 7 | 100 | 0 |
| 8 | 20 | 0 |
| 9 | 68 | 1 |
| 10 | 70 | 1 |
| 11 | 90 | 1 |
| 12 | 45 | 1 |
| 13 | 75 | 1 |
| 14 | 80 | 1 |
| 15 | 50 | 1 |
| 16 | 60 | 0 |
| 17 | 85 | 1 |
| 18 | 55 | 0 |
| 19 | 75 | 1 |
| 20 | 68 | 1 |
| 21 | 44 | 0 |
| 22 | 80 | 1 |
| 23 | 65 | 1 |
| 24 | 90 | 1 |
| 25 | 110 | 1 |
| 26 | 35 | 0 |
| 27 | 100 | 1 |
| 28 | 60 | 1 |
| 29 | 55 | 1 |
| 30 | 50 | 0 |
| 31 | 60 | 0 |
| 32 | 75 | 1 |
| 33 | 100 | 1 |
| 34 | 90 | 1 |
| 35 | 75 | 1 |
| 36 | 60 | 1 |
| 37 | 60 | 1 |
| 38 | 50 | 1 |
| 39 | 95 | 1 |
| 40 | 74 | 0 |
| 41 | 55 | 1 |
| 42 | 120 | 1 |
| 43 | 60 | 1 |
| 44 | 80 | 1 |
| 45 | 70 | 1 |
| 46 | 85 | 1 |
| 47 | 110 | 1 |
| 48 | 65 | 1 |
| 49 | 95 | 1 |
| 50 | 55 | 0 |

Для нахождения неизвестных параметров  будем находить максимум по этим параметрам функции (2). Сначала, используя метод случайного поиска, найдем .

Параметры случайного поиска:

,

область изменения параметров .

Таблица 4.2. Результаты случайного поиска

|  |  |  |
| --- | --- | --- |
|  |  |  |
| -4,24 | 0,0899 | -21,6028 |

Как правило, случайный поиск экстремальной точки играет роль поиска области, в которой находится глобальный экстремум, то есть играет роль “разведки”. Для уточнения решения задачи далее следует применить один из направленных методов. Таким образом, случайный поиск более эффективен в сочетании с детерминированными направленными методами. Рассмотрим один из них.

**Классический градиентный метод.**

Рассмотрим подробнее классический градиентный метод. Основная идея этого метода заключается в том, что для поиска максимума целевой функции достаточно идти в направлении ее градиента

, (4.3)

где шаг градиентного метода. Для частного случая вектор градиента для целевой функции (4.2) имеет вид

. (4.4)

Запишем (4.3) по переменным

,

. (4.5)

В качестве начальной точки использовался результат случайного поиска: , начальный шаг равен и .

В таблице 4.3 приводятся значения параметров и целевой функции для ряда итераций.

Таблица 4.3. Результаты уточнения найденных коэффициентов при использовании градиентного метода

|  |  |  |  |
| --- | --- | --- | --- |
| № |  |  |  |
| 1 | -4,24012215133595 | 0,0807315071670666 | -21,6024474500769 |
| 2 | -4,24008189301828 | 0,0813794122165656 | -21,5351977815969 |
| 3 | -4,23998911250237 | 0,0861350087392071 | -21,3644406819876 |
| 4 | -4,23998962350153 | 0,0851202782453369 | -21,3555022065666 |
| 5 | -4,23997122404936 | 0,0852607349546962 | -21,3553299462322 |
| 6 | -4,23995547952403 | 0,0852385771289727 | -21,3553206985405 |
| 7 | -4,23993932289043 | 0,0852416702651814 | -21,3553154164727 |
| 8 | -4,23992322974969 | 0,085240891680852 | -21,3553102276764 |
| 9 | -4,23844302753144 | 0,0852167886709811 | -21,3548338011444 |

В результате получили максимум логарифма функции правдоподобия

,

который достигается при следующих коэффициентах регрессии:

.

Зная значения оптимальных параметров, получаем модель логистической регрессии для данной задачи:

. (4.6)

Найдем точку деления по зарплате. При функция (граница разделения). Следовательно, если , то является точкой деления исходных данных на две группы

т.р. (4.7)

Точка x=49,74 т.р. является граничной точкой в данной одномерной задаче.

Рис.4.1. Зависимость платёжеспособности и зарплаты

В таблице 4.4 приведены ***результаты прогнозирования одномерной логит-модели по следующему правилу***.



Таблица 4.4. Сравнение результатов модели с исходными данными

и прогноз для 10 клиентов в конце таблицы

|  |  |  |  |
| --- | --- | --- | --- |
| **Данные о заемщиках** | | | **Выходные данные модели** |
| **№** | **Зарплата, т.р.** | **Выдача/отказ** | **вероятность** |
| 1 | 55 | 0 | 0,610231 |
| 2 | 40 | 0 | 0,303534 |
| 3 | 75 | 1 | 0,895982 |
| 4 | 55 | 1 | 0,610231 |
| 5 | 45 | 1 | 0,40029 |
| 6 | 40 | 0 | 0,303534 |
| 7 | 100 | 0 | 0,98641 |
| 8 | 20 | 0 | 0,0734 |
| 9 | 68 | 1 | 0,825862 |
| 10 | 70 | 1 | 0,84904 |
| 11 | 90 | 1 | 0,968695 |
| 12 | 45 | 1 | 0,40029 |
| 13 | 75 | 1 | 0,895982 |
| 14 | 80 | 1 | 0,929539 |
| 15 | 50 | 1 | 0,505503 |
| 16 | 60 | 0 | 0,705692 |
| 17 | 85 | 1 | 0,95284 |
| 18 | 55 | 0 | 0,610231 |
| 19 | 75 | 1 | 0,895982 |
| 20 | 68 | 1 | 0,825862 |
| 21 | 44 | 0 | 0,380009 |
| **22** | 80 | 1 | 0,929539 |
| **23** | 65 | 1 | 0,785973 |
| **24** | 90 | 1 | 0,968695 |
| **25** | 110 | 1 | 0,99416 |
| **26** | 35 | 0 | 0,221526 |
| **27** | 100 | 1 | 0,98641 |
| **28** | 60 | 1 | 0,705692 |
| **29** | 55 | 1 | 0,610231 |
| **30** | 50 | 0 | 0,505503 |
| **31** | 60 | 0 | 0,705692 |
| **32** | 75 | 1 | 0,895982 |
| **33** | 100 | 1 | 0,98641 |
| **34** | 90 | 1 | 0,968695 |
| **35** | 75 | 1 | 0,895982 |
| **36** | 60 | 1 | 0,705692 |
| **37** | 60 | 1 | 0,705692 |
| **38** | 50 | 1 | 0,505503 |
| **39** | 95 | 1 | 0,979335 |
| **40** | 74 | 0 | 0,887764 |
| **41** | 55 | 1 | 0,610231 |
| **42** | 120 | 1 | 0,997502 |
| **43** | 60 | 1 | 0,705692 |
| **44** | 80 | 1 | 0,929539 |
| **45** | 70 | 1 | 0,84904 |
| **46** | 85 | 1 | 0,95284 |
| **47** | 110 | 1 | 0,99416 |
| **48** | 65 | 1 | 0,785973 |
| **49** | 95 | 1 | 0,979335 |
| **50** | 55 | 0 | 0,610231 |
| **51** | 57 |  | 0,649943 |
| **52** | 68 |  | 0,825862 |
| **53** | 73 |  | 0,878985 |
| **54** | 86 |  | 0,956526 |
| **55** | 46 |  | 0,420919 |
| **56** | 64,5 |  | 0,778715 |
| **57** | 53 |  | 0,569001 |
| **58** | 98 |  | 0,983924 |
| **59** | 55 |  | 0,610231 |
| **60** | 49,5 |  | 0,494847 |

Находим по логистической модели значение вероятности. Построим таблицу количества истинных и ложных прогнозов:

Таблица 4.5. Матрица классов, соответствующая порогу 0.5

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Количество |
|  | 35(*TP*) | 2(*FN)* | 37 |
|  | 8(*FP*) | 5(*TN*) | 13 |
| Количество | 43 | 7 | 50 |

Как видно из таблицы 4.5 модель допускает 2 ошибок II рода и 8 ошибок I рода на данной обучающей выборке.

**Ошибкой I рода** является случай, когда по прогнозу клиент не должен был получить кредит, но получил. В таком случае банк рискует, что кредит может быть не выплачен, что приведет к убыткам.

**Ошибкой II** рода является случай, когда прогноз для клиента положительный, но кредит не выдается. Потенциально банк упускает возможность заработать на процентах с кредита.

**ROC-анализ**

Применим ROC-анализ для для нахождения оптимального порога отсечения. Для этого построим sensitivity и specificity в зависимости от cutoff, который меняется в пределах от 0 до 1 по формулам (3.9) и (3.10) соответственно.

На рис. 4.1 кривая оранжевого цвета отражает изменение значения sensitivity в зависимости от cutoff, а кривая зеленого цвета – значения specificity.

Таблица 4.6. Зависимость чувствительности и специфичности от порога.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cutoff | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
| sensitivity | 1 | 1 | 1 | 1 | 0,94 | 0,89 | 0,81 | 0,64 | 0.40 |
| specificity | 0,076 | 0,076 | 0,15 | 0,38 | 0,38 | 0,46 | 0,69 | 0,84 | 0,92 |

Построив на одном графике линии sensitivity и specificity, можем определить оптимальный порог в точке их пересечения.

Рис. 4.2. График пересечения чувствительности и специфичности 1 предиктора

Оптимальный cutoff =0,688.

Общий показатель успеха (Оverall success rate) представляет собой число правильно классифицированных наблюдений, отнесенное к общему числу наблюдений:

. (4.8)

Таким образом точность модели составила 74%.

**Вывод.** В таблице 4.4 последние 10 клиентов отобраны для прогноза решения о выдаче кредита. Если cutoff =0,5, то клиентам под номером 55 и 60 кредит не выдадут, а при оптимальном cutoff =0,688 в категорию клиентов, которые не получат кредит, попадут клиенты под номерами 51, 55, 57, 59 и 60.

**Задача 2. Логистическая модель с 6 предикторами**

Изначально предполагалось, что все предикторы из выборки влияют на решение банка о выдаче кредита или отказе в нем, поэтому составим модель с учетом всех предикторов:

* + - 1. Заработная плата (т.р.)
      2. Возраст (лет)
      3. Дети
      4. Образование
      5. Рабочая сфера
      6. Пол

**Постановка задачи**

Используя логистическую регрессию, нужно определить вероятность выдачи кредита заемщику, зная 6 его характеристик (предикторов). В таком случае логарифм функции максимального правдоподобия будет иметь такой вид:

(4.9)

Для нахождения глобального максимума снова сначала применим простейший случайный поиск и найдем оптимальные параметры логит-модели .:

Параметры случайного поиска:

Границы изменения

Таблица 4.7. Результаты случайного поиска для 6 предикторов

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| -0,14 | 0,17 | 0,65 | 3,55 | -064 | -0,38 | 0,2 | -487,46 |

Уточним результаты с помощью градиентного метода.

В качестве начальной точки использовался результат случайного поиска: и возьмем начальный шаг: h=0,0001.

В таблице 4.8 приведены результаты итераций спуска при применении градиентного метода:

Таблица 4.8. Результаты спуска при 6 предикторах

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| -0,145 | 0,183 | -0,335 | 3,2 | -0,55 | 0,3 | 0,25 | -52,195106 |

Тогда функция логистический регрессии в данной задаче примет следующий вид:

(4.10)

В таблице 4.9 приведено сравнение результатов модели с исходными данными:

Таблица 4.9. Сравнение исходных данных с результатами модели для 6 предикторов

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **№** | **ЗП, т.р.** | **Возраст** | **Дети** | **Образование** | **Рабочая сфера** | **Пол** | **Отклик** | **Вероятность** |
| 1 | 55 | 39 | 1 | 1 | 1 | 1 | 0 | 0,326192 |
| 2 | 40 | 56 | 1 | 3 | 2 | 2 | 0 | 0,000427 |
| 3 | 75 | 27 | 2 | 2 | 1 | 2 | 1 | 0,999959 |
| 4 | 55 | 43 | 1 | 1 | 1 | 2 | 1 | 0,130539 |
| 5 | 45 | 29 | 0 | 1 | 1 | 2 | 1 | 0,123875 |
| 6 | 40 | 22 | 1 | 1 | 1 | 1 | 0 | 0,298682 |
| 7 | 100 | 46 | 3 | 3 | 1 | 1 | 0 | 0,999967 |
| 8 | 20 | 61 | 0 | 1 | 1 | 2 | 0 | 0, 168\* |
| 9 | 68 | 47 | 2 | 1 | 1 | 2 | 1 | 0,905973 |
| 10 | 70 | 41 | 0 | 3 | 2 | 2 | 1 | 0,080059 |
| 11 | 90 | 40 | 1 | 3 | 1 | 2 | 1 | 0,988769 |
| 12 | 45 | 27 | 1 | 1 | 1 | 1 | 1 | 0,846327 |
| 13 | 75 | 27 | 0 | 2 | 1 | 1 | 1 | 0,969147 |
| 14 | 80 | 26 | 0 | 3 | 1 | 2 | 1 | 0,988083 |
| 15 | 50 | 29 | 0 | 1 | 1 | 1 | 1 | 0,215715 |
| 16 | 60 | 30 | 2 | 2 | 1 | 1 | 0 | 0,997621 |
| 17 | 85 | 37 | 3 | 3 | 2 | 2 | 1 | 0,999988 |
| 18 | 55 | 39 | 2 | 1 | 1 | 2 | 0 | 0,338585 |
| 19 | 75 | 39 | 2 | 1 | 1 | 1 | 1 | 0,997847 |
| 20 | 68 | 31 | 0 | 3 | 2 | 2 | 1 | 0,677973 |
| 21 | 44 | 25 | 0 | 1 | 1 | 2 | 0 | 0,327734 |
| 22 | 80 | 32 | 0 | 2 | 1 | 2 | 1 | 0,944632 |
| 23 | 65 | 30 | 1 | 2 | 2 | 2 | 1 | 0,986649 |
| 24 | 90 | 31 | 2 | 3 | 1 | 2 | 1 | 0,999981 |
| 25 | 110 | 40 | 3 | 3 | 1 | 2 | 1 | 0,999992 |
| 26 | 35 | 19 | 0 | 1 | 1 | 1 | 0 | 0,381 |
| 27 | 100 | 42 | 2 | 2 | 1 | 2 | 1 | 0,999913 |
| 28 | 60 | 26 | 1 | 2 | 1 | 1 | 1 | 0,986043 |
| 29 | 55 | 23 | 1 | 2 | 2 | 2 | 1 | 0,993023 |
| 30 | 50 | 34 | 2 | 1 | 1 | 2 | 0 | 0,973071 |
| 31 | 60 | 48 | 2 | 1 | 1 | 2 | 0 | 0,609517 |
| 32 | 75 | 41 | 2 | 2 | 2 | 2 | 1 | 0,995629 |
| 33 | 100 | 46 | 1 | 3 | 1 | 2 | 1 | 0,984897 |
| 34 | 90 | 34 | 2 | 3 | 1 | 2 | 1 | 0,999945 |
| 35 | 75 | 37 | 1 | 2 | 1 | 2 | 1 | 0,96601 |
| 36 | 60 | 29 | 0 | 2 | 1 | 1 | 1 | 0,497498 |
| 37 | 60 | 25 | 0 | 3 | 1 | 1 | 1 | 0,702774 |
| 38 | 50 | 31 | 0 | 2 | 2 | 2 | 1 | 0,11907 |
| 39 | 95 | 37 | 3 | 3 | 2 | 2 | 1 | 0,999998 |
| 40 | 74 | 43 | 2 | 2 | 1 | 1 | 0 | 0,981738 |
| 41 | 55 | 36 | 0 | 1 | 1 | 2 | 1 | 0,068392 |
| 42 | 120 | 37 | 3 | 3 | 1 | 2 | 1 | 0,999996 |
| 43 | 60 | 32 | 1 | 2 | 2 | 2 | 1 | 0,935637 |
| 44 | 80 | 30 | 1 | 2 | 1 | 1 | 1 | 0,998497 |
| 45 | 70 | 43 | 0 | 2 | 2 | 1 | 1 | 0,054588 |
| 46 | 85 | 28 | 2 | 3 | 1 | 2 | 1 | 0,999984 |
| 47 | 110 | 45 | 2 | 2 | 1 | 1 | 1 | 0,999948 |
| 48 | 65 | 27 | 0 | 2 | 1 | 2 | 1 | 0,866015 |
| 49 | 95 | 36 | 1 | 3 | 1 | 2 | 1 | 0,998903 |
| 50 | 55 | 39 | 1 | 1 | 1 | 1 | 0 | 0,326192 |

\*комментарии к таблице:

Рабочая сфера: 1- частная, 2 – гос. Служащие

Образование: 1- СПО, 2 – бакалавриат, 3 – магистратура,

Пол: 1 – женский, 2 – мужской.

Далее построим четырехпольную таблицу, чтобы определить количество истинных и ложных прогнозов, порогом снова возьмем cutoff = 0,5.

Таблица 4.10. Матрица классов для cutoff = 0,5

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Количество |
|  | 29(*TP*) | 8(*FN)* | 37 |
|  | 5(*FP*) | 8(*TN*) | 13 |
| Количество | 34 | 16 | 50 |

Проведем ROC-анализ:

Таблица 4.11. Зависимость чувствительности и специфичности от порога.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cutoff | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
| Sensitivity | 0,91 | 0,88 | 0,81 | 0,81 | 0,783 | 0,783 | 0,756 | 0,729 | 0,67 |
| Specificity | 0,15 | 0,16 | 0,23 | 0,61 | 0,62 | 0,62 | 0,69 | 0,69 | 0,69 |

По графику определим оптимальный порог:

Рис. 4.3 Пересечение чувствительности и специфичности для 6 предикторов

Оптимальный cutoff = 0,8.

Общий показатель успеха (Оverall success rate)

**Задача 3. Логистическая модель для оптимального количества предикторов**

Для начала определим значимость коэффициентов, для этого вычислим значение статистики ** по формуле 2.13.

В таблице 4.12 приведены результаты расчета статистических гипотез:

Таблица 4.12. Результат статистических гипотез

|  |  |
| --- | --- |
|  | **Z** |
|  | 23,38519882 |
|  | -28,8107619 |
|  | 273,0895028 |
|  | -2,86666182 |
|  | 1,093149687 |
|  | 1,124178604 |

**Вывод.** После проведения проверки статистических гипотез было установлено, что коэффициент при предикторе («рабочая сфера» и «пол») является незначимым.

**Постановка задачи**

Требуется построить логит-модель для оптимального количества предикторов (m=4) и, используя логистическую регрессию, определить вероятность выдачи кредита заемщику.

В таком случае логарифм функции максимального правдоподобия будет иметь такой вид:

(4.15)

Для нахождения глобального максимума снова применим простейший случайный поиск и найдем . На первом шаге получим значение параметров :

Параметры случайного поиска:

N  = 50,

Границы изменения предикторов (-10;10), .

Таблица 4.13. Результаты случайного поиска для 4 предикторов

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| -0,95 | 0,218 | -0,438 | -0,188 | 0,696 | -42,998 |

Уточним результаты опять с помощью градиентного метода.

В качестве начальной точки использовался результат случайного поиска:

Первоначальный шаг: h=0,0001

В таблице 4.14 приведены результаты итераций спуска при применении градиентного метода:

Таблица 4.14. Результаты спуска при 4 предикторах

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| -0,95 | 0,099 | -0,14 | -0,201 | 0,73 | -18,288 |

Тогда функция логистический регрессии в данной задаче примет следующий вид:

(4.16)

В таблице 4.15 приведено сравнение результатов модели с исходными данными:

Таблица 4.15. Результаты прогнозирования

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№** | **ЗП, т.р.** | **Возраст** | **Дети** | **Образование** | **Отклик** | **Вероятность** |
| 1 | 55 | 39 | 1 | 1 | 0 | 0,392867448 |
| 2 | 40 | 56 | 1 | 3 | 0 | 0,055107021 |
| 3 | 75 | 27 | 2 | 2 | 1 | 0,977187562 |
| 4 | 55 | 43 | 1 | 1 | 1 | 0,269801735 |
| 5 | 45 | 29 | 0 | 1 | 1 | 0,543666121 |
| 6 | 40 | 22 | 1 | 1 | 0 | 0,613083128 |
| 7 | 100 | 46 | 3 | 3 | 0 | 0,983738995 |
| 8 | 20 | 61 | 0 | 1 | 0 | 0,001130139 |
| 9 | 68 | 47 | 2 | 1 | 1 | 0,385019703 |
| 10 | 70 | 41 | 0 | 3 | 1 | 0,919200647 |
| 11 | 90 | 40 | 1 | 3 | 1 | 0,987292183 |
| 12 | 45 | 27 | 1 | 1 | 1 | 0,56344642 |
| 13 | 75 | 27 | 0 | 2 | 1 | 0,984596323 |
| 14 | 80 | 26 | 0 | 3 | 1 | 0,996023972 |
| 15 | 50 | 29 | 0 | 1 | 1 | 0,66159795 |
| 16 | 60 | 30 | 2 | 2 | 0 | 0,864264296 |
| 17 | 85 | 37 | 3 | 3 | 1 | 0,979716362 |
| 18 | 55 | 39 | 2 | 1 | 0 | 0,346286833 |
| 19 | 75 | 39 | 2 | 1 | 1 | 0,793450306 |
| 20 | 68 | 31 | 0 | 3 | 1 | 0,974276621 |
| 21 | 44 | 25 | 0 | 1 | 0 | 0,653939647 |
| 22 | 80 | 32 | 0 | 2 | 1 | 0,981155242 |
| 23 | 65 | 30 | 1 | 2 | 1 | 0,927345315 |
| 24 | 90 | 31 | 2 | 3 | 1 | 0,995563391 |
| 25 | 110 | 40 | 3 | 3 | 1 | 0,997358457 |
| 26 | 35 | 19 | 0 | 1 | 0 | 0,642302025 |
| 27 | 100 | 42 | 2 | 2 | 1 | 0,984212288 |
| 28 | 60 | 26 | 1 | 2 | 1 | 0,93160755 |
| 29 | 55 | 23 | 1 | 2 | 1 | 0,92666824 |
| 30 | 50 | 34 | 2 | 1 | 0 | 0,394061452 |
| 31 | 60 | 48 | 2 | 1 | 0 | 0,197675771 |
| 32 | 75 | 41 | 2 | 2 | 1 | 0,857677705 |
| 33 | 100 | 46 | 1 | 3 | 1 | 0,989043912 |
| 34 | 90 | 34 | 2 | 3 | 1 | 0,993261476 |
| 35 | 75 | 37 | 1 | 2 | 1 | 0,928016625 |
| 36 | 60 | 29 | 0 | 2 | 1 | 0,916177618 |
| 37 | 60 | 25 | 0 | 3 | 1 | 0,975453311 |
| 38 | 50 | 31 | 0 | 2 | 1 | 0,754119965 |
| 39 | 95 | 37 | 3 | 3 | 1 | 0,992370499 |
| 40 | 74 | 43 | 2 | 2 | 0 | 0,804851148 |
| 41 | 55 | 36 | 0 | 1 | 1 | 0,546147517 |
| 42 | 120 | 37 | 3 | 3 | 1 | 0,99935437 |
| 43 | 60 | 32 | 1 | 2 | 1 | 0,854596703 |
| 44 | 80 | 30 | 1 | 2 | 1 | 0,982579721 |
| 45 | 70 | 43 | 0 | 2 | 1 | 0,805479042 |
| 46 | 85 | 28 | 2 | 3 | 1 | 0,995219276 |
| 47 | 110 | 45 | 2 | 2 | 1 | 0,991013189 |
| 48 | 65 | 27 | 0 | 2 | 1 | 0,959573364 |
| 49 | 95 | 36 | 1 | 3 | 1 | 0,995541238 |
| 50 | 55 | 39 | 1 | 1 | 0 | 0,392867448 |
| Прогноз для 10 новых клиентов | | | | | | |
| 51 | 40 | 25 | 2 | 2 |  | 0,638846279 |
| 52 | 60 | 33 | 1 | 1 |  | 0,711066355 |
| 53 | 53 | 30 | 3 | 1 |  | 0,556542961 |
| 54 | 60 | 38 | 1 | 3 |  | 0,86543388 |
| 55 | 38 | 21 | 0 | 3 |  | 0,887284921 |
| 56 | 66 | 40 | 3 | 2 |  | 0,699419123 |
| 57 | 70 | 45 | 0 | 3 |  | 0,866594941 |
| 58 | 81 | 50 | 2 | 2 |  | 0,755786003 |
| 59 | 50 | 28 | 1 | 1 |  | 0,648028723 |
| 60 | 77 | 47 | 2 | 2 |  | 0,760191294 |

Проведем ROC-анализ для оптимальной модели с 4 предикторами, для начала построим четырехпольную таблицу:

Таблица 4.16. Матрица классов для 4-х предикторов, порог отсечения =0,5

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Количество |
|  | 32(*TP*) | 5(*FN)* | 37 |
|  | 6(*FP*) | 7(*TN*) | 13 |
| Количество | 38 | 12 | 50 |

Найдем зависимости чувствительности и специфичности от порога, построим график для определения оптимального порога:

Таблица 4.17. Зависимость чувствительности и специфичности от выбранного порога.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cutoff | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
| sensitivity | 1 | 1 | 0,972 | 0,945 | 0,945 | 0,842 | 0,837 | 0,783 | 0,702 |
| specificity | 0,153 | 0,23 | 0,23 | 0,538 | 0,538 | 0,583 | 0,769 | 0,769 | 0,923 |

График пересечения специфичности и чувствительности:

Рис. 4.4. Пересечение sensitivity и specificity при 4 предикторах

Оптимальный cutoff = 0,857

Общий показатель успеха (Оverall success rate) по формуле 2.8:

**Вывод.** В таблице 4.15 последние 10 клиентов отобраны для прогнозирования решения о выдаче кредита. При оптимальном cutoff=0.857 в категория клиентов, получивших кредит, попадут клиенты с номерами 54,55,57.

**Построение доверительных интервалов**

1. Формула для доверительного интервала с уровнем доверия

()

для любого параметра имеет вид:

, (4.17)

где – квантиль уровня распределения Стьюдента :

,

где – плотность распределения вероятностей Стьюдента. (Так обозначена функция. Не путайте с обозначением величины S перед корнем в интервальной оценке). Величина находится по таблице Стьюдента по данным

и ,

1. Матрица размерности для предикторов и клиентов

(4.18)

и матрица равна

.

Это симметричная матрица относительно диагонали размерности .

1. Найти элементы обратной матрицы . Она тоже квадратная и размерности . Величины являются ее диагональными элементами.

Предполагаем нормально распределенными, независимыми, центрированными и некоррелированными:

, , (4.19)

.

– независимые равномерно распределенные числа в (0,1), .

Результаты интервальных оценок:

Для коэффициентов логистической регрессии, найденных методом максимального правдоподобия, построены доверительные интервалы с уровнем доверия 0.95. Найденные интервалы покрывают истинные значения коэффициентов с заданной вероятностью.

**Выводы**

При решении задачи на основе обучающей выборки о выдаче кредита были построены 3 модели.

1. Модель первая – одномерная. Использовался один параметр – зарплата. На основе модели была визуализирована сигмоид-кривая, найдена граничная точка и функция максимального правдоподобия. Граничная точка модели соответствует минимальному значению зарплаты для выдачи кредита, она равна 49,74 т.р. Точность модели составила 74%, что классифицирует ее как хорошую.
2. Вторая модель в работе была построена с учетом всех имеющихся 6 предикторов. Целью ее исследования было выявление статистически незначимых предикторов при использовании гипотезы о значимости параметров. Ими оказались: пол и рабочая сфера. Так как влияние этих данных незначительно, их можно исключить из модели. Точность модели при всех имеющихся предикторах составила 72%, что характеризует модель как хорошую.
3. Третья модель была построена на основе оптимального количества предикторов, которых было 4: зарплата, возраст, образование и дети. Ниже в таблице сопоставлены точности моделей.

Таблица 4.16. Сравнение логит-моделей с разным количеством предикторов

|  |  |  |
| --- | --- | --- |
| **Модель** | **Точность** | **Количество ошибок** |
| Одномерная модель (зарплата) | 74% | 11 |
| Модель со всеми известными предикторами | 72% | 14 |
| Модель с оптимальным количеством предикторов | 76% | 12 |

Как видно из таблицы, добавление предикторов влияет на качество логит-моделей. При всех известных предикторах качество модели самое низкое. При получении оптимального значения предикторов, мы имеем самую высокую точность модели. Самый весомый вклад в прогноз вносят заработная плата и возраст.

ROC-анализ позволил определить оптимальный порог отсечения, который составил 0,857. Таким образом, общая точность модели возросла на 4 %, чувствительность не изменилась, а специфичность возросла 18,44%, удалось избежать двух ошибок 1 рода, что позволит банку избежать убытков.

Из приведенного выше можно заключить, что ***логит-модель на основе обучающей выборки с оптимальным количеством предикторов является наилучшей.*** Оптимальными характеристиками стали зарплата, возраст, образование и дети, порог отсечения равен 0,857.

# Заключение.

В нынешних условиях, когда количество обращений за кредитом растет, банкам не обойтись без частичной автоматизации процесса принятия решения о выдаче. Информационные технологии расширяют возможности статистического анализа, позволяют собирать данные, анализировать их, автоматизировать и оптимизировать различные процессы.

Логистическая регрессия является хорошим способом аналитики данных о заемщиках, позволяет на основе характеристик потенциального должника автоматически определять его благонадежность.

В данной работе была построена оптимальная логит-модель, которая определяла вероятность выдачи кредита на основе обучающей выборки.

Эта сложная задача решалась в 3 этапа. Первой простейшей задачей являлась задача с одним предиктором. Она наглядно показала суть задачи. Однако она далека от жизненной ситуации, когда можно принимать решение только на основании величины заработной платы.

Поэтому далее перешли ко второй задаче с 6 предикторами.

Благодаря проверке гипотез о значимости коэффициентов, было получено оптимальное число параметров и решена третья задача с этим числом параметров.

Для решения этих задач

* использовался метод максимального правдоподобия,
* написана программа для ЭВМ на языке (C#) для поиска максимума функции правдоподобия, зависящей от параметра (а именно 2, 7, 5),
* оптимизация проводилась на основе применения случайного поиска в сочетании с градиентным методом,
* на основе проверки гипотез о значимости параметров логит-модели выделены значимые предикторы, их оказалось 4 вместо 6,
* для оптимального числа параметров рассчитаны доверительные интервалы для каждого коэффициенты логистической регрессии,
* для оценки качества логистической модели применялся ROC-анализ:

1) cтроились кривые чувствительности и специфичности модели, которые показывают зависимость верно классифицированных положительных и отрицательных примеров от порога отсечения,

2) выявили оптимальный порог отсечения.

Компьютерные расчеты, анализ, построение диаграмм и графиков были проведены с помощью написанных ЭВМ (C#) и Excel.

Обучающая выборка составлялась с учетом данных [15].

# Список литературы

1. Буре, В. М. Методы прикладной статистики в R и Excel: учебник для вузов/ Парилина, Е. М., Седаков, А.А. – Санкт-Петербург: Лань, 2016.
2. Ван дер Варден Б. Л. Математическая статистика. – М.: ИЛ, 1960.
3. Владимирова Л.В., Овсянников Д.А., Рубцова И.Д. Методы Монте-Карло в прикладных задачах. СПб.: Изд-во ВВМ, 2015. 167 с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика. М. “Высшая школа”, 1998, 479 с.
5. Ермаков С.М. Методы Монте-Карло+ и смежные вопросы. Изд-во “Наука”, М. 1975, 472 с.
6. Ермаков С.М. Математическая теория планирования эксперимента. М.: Наука, 1983. – 392 с.
7. Карманов В.Г. Математическое программирование. М. ”Наука”, 1986, 288 с.
8. Паклин Н.Б. Логистическая регрессия и ROC-анализ – математический аппарат // Официальный сайт компании BaseGroup Labs URL: https://basegroup.ru/community/articles/logistic.
9. Сорокин А.С. Построение скоринговых карт с использованием модели логистической регрессии / А.С. Сорокин // Интернет-журнал «НАУКОВЕДЕНИЕ». – 2014. -№2.
10. Churchill G. A., Nevin J. R., Watson R. R.//The role of credit scoring in the loan decision. Credit World. March/1977
11. Greene W.H. Econometric Analysis, 5th edition, New Jearsey: Pearson Education, 2003.
12. Hand D. J., Henley W.E. Statistical classification methods in consumer credit // Journal of the Royal Statistical Society, 1997. P. 532 – 541.
13. Hosmer D., Lemeshow S. Applied logistic regression. N. Y.: Wiley, 2000. 375 p.

Myers J. H., Forgy E. W. The development of numerical credit evaluation systems//Journal of American Statistical Association. September/1963

Используемые интернет ресурсы:

<https://archive.ics.uci.edu/ml/machine-learning-databases/adult>