УДК 004.942

Разработка скоринговой модели с использованием методов логистической регрессии и ROC – анализа

И.Ю. Анохина
Донецкий национальный технический университет
ingatula@mailru

Анохина И.Ю. Разработка скоринговой модели м использованием методов логистической регрессии и ROC – анализа. *Рассмотрены вопросы создания скоринговой модели на основании данных медицинских исследований. При моделировании использованы методы интеллектуального анализа данных, к которым относится аппарат логистического регрессионного анализа, логико-статистические подходы к распознаванию образов, в том числе ROC – анализ. Описаны этапы разработки скоринговой модели. Проведена оценка точности моделирования на основании статистических данных, полученных при клинических испытаниях.*

Ключевые слова: *статистические данные, скоринговые модели, методология разработки, риски, логистическая регрессия, ROC – анализ, пакет Statistica.*

Скоринг и его история

Скоринг (от англ. scoring – подсчет очков в игре) - модель классификации статистической базы данных по различным группам, если неизвестна характеристика, которая разделяет эти группы, но имеется набор параметров, оказывающих влияние на исследуемую характеристику.

Классификация исследуемого объекта осуществляется на основании скоринговой карты, с помощью которой рассчитывается скоринговый балл конкретного человека.

Скоринговые модели и карты получили широкое распространение в различных областях жизнедеятельности человека.

Например, в банковской сфере это кредитный скоринг. По социальным характеристикам клиента (пол, возраст, место проживания, должность и т.д.) можно, на основе его анкеты, отнести клиента к группе сильно или слабо соответствующих бизнесу. Исходя из рассчитанного уровня доверия, принимается решение о предоставлении кредита. Основной целью скоринга является классификация клиентов банка на «хороших» и «плохих», исходя из чего кредитор может выбирать соответствующие действия по отношению к клиенту. Под «плохим» клиента, понимают клиента с низкой прогнозируемой вероятностью возвращения кредита.

Как раз с кредитного скоринга, видимо, и началась история его использования. В 1941 Давид Дюран опубликовал первую исследовательскую работу по кредитному скорингу, в которой оценивал роль различных факторов на формирование рейтинга клиента. Взрыв спроса на кредиты после окончания Второй мировой войны обусловил необходимость автоматизации процесса принятия решений о выдаче кредитов. В 60-х годах началось внедрение компьютерных технологий в область скоринга.

Страховым компаниям необходимо принимать решение, стоит ли страховать конкретного человека и какую сумму он должен платить за страховку. При этом они основываются на вероятности наступления страховых случаев, например, аварий, если рассматривается вопрос автомобильного страхования. Факторами принятия решений являются такие как: история нарушений, опыт вождения, год приобретения и состояние автомобиля и т. д. Анализируя их, с помощью скоринговых карт клиента, прогнозируют риски.

Проведенные исследования, показали, что кредитная история является неоспоримым и точным индикатором риска того, что с данным лицом произойдет страховой случай. Для многих страхователей скоринговые системы позволяют платить меньшие страховые премии.

В фармацевтике скоринговые модели используются для определение класса препаратов, которые будут эффективны для различных групп клиентов.

В молекулярной генетике – это молекулярный диагноз заболеваний с учетом того, что генетический код можно использовать для расчета вероятности заболевания.

Постановка задачи

В современной медицине существует ряд задач, для решения которых требуются количественные, оценочные методы.

Поэтому существует необходимость в разработке математических моделей описания процессов, свойств человеческого организма и создания на их базе скоринговых карт или систем, описывающих состояние пациента, степень нарушения функций организма, степень тяжести его состояния.

Скоринговые системы позволяют определять вероятность летального исхода или выздоровления пациента, оценивать степень и стадии заболевания, причем в виде числовых характеристик, что невозможно даже для врача-клинициста с большим опытом. Такая система строится на основе модели, учитывающей симптомы заболевания и само заболевание [1].

Существуют два основных метода создания скоринговых моделей. Первый основан на анализе статистических данных, второй на экспертной оценке. Основная проблема первого метода - получение репрезентативной базы данных, содержащей не только часто встречающиеся, но и редкие случаи проявления болезней.

Второй метод (экспертные системы) - это интеллектуальные системы, ориентированные на тиражирование опыта высококвалифицированных специалистов (экспертов) в таких областях медицины, в которых качество принятия решений зависит от уровня экспертизы. Основная проблема этого метода - трудоемкость выявления факторов, сложность разработки и неполнота множества правил.

Чаще в медицине используют первый метод, основанный на анализе статистических данных.

Нам были предоставлены статистические данные, полученные путем анкетирования пациентов, проводимого в медицинских учреждениях г. Донецка. Поставлена задача разработки скоринговой модели для определения степени риска возникновения заболевания микрохолелитиаза (далее МХЛ, в просторечии, образование желудочных камней или сладжа).

Как показали исследования, существует ряд факторов, влияющих на вероятность образования сладжа. Частично эти факторы определяются путем анкетирования пациента. Учитывая, что клинические испытания достаточно дороги, целесообразно на первом этапе исследований выявить группы лиц, предрасположенных к заболеванию, и только в случае высокой степени риска назначать клинические обследования.

Была обследована группа из 131 пациента с установленным заболеванием и контрольная группа обследуемых, у которых диагноз не подтвердился.

Моделирование проводили по следующему алгоритму:

1. Создание скоринговой карты (перечня учитываемых в модели параметров, а именно факторов, способствующих возникновению заболевания).
2. Выбор метода обработки статистических данных.
3. Классификация обследуемых, т.е. создание модели на основе части выборки. Результатом этого этапа становится скоринговая модель, согласно которой по имеющейся информации о конкретном пациенте, выдается оценка степени риска заболевания в баллах.
4. Проверка модели на оставшейся части выборки.
5. При необходимости доработка модели с учетом новых статистических данных.

Первоначально исследовалась выборка, в которой количество анализируемых факторов было более двадцати. Проведение факторного анализа позволило сократить число исследуемых переменных до шести. Как показал анализ факторных нагрузок, на увеличение вероятности риска заболевания существенное влияние оказывают: наличие травм в анамнезе, индекс массы тела (здесь и далее ИМТ), уровень физической активности пациента, возраст, наследственность и изменения в режиме питания, включающие в себя соблюдение постов и голодание сроком более семи, а особенно более четырнадцати дней.

Для решения поставленной задачи выбран метод логистической регрессии.

Создание скоринговой модели

Логистическая регрессия – разновидность множественной регрессии, позволяющая установить зависимость между несколькими независимыми переменными и зависимой, причем как зависимая, так и/или независимые переменные могут принимать два значения (да/нет), в дальнейшем будем называть их бинарными.

Задачи, в которых используются бинарные коэффициенты, встречаются достаточно часто. Например, в качестве бинарной переменной можно рассматривать переменную, характеризующую отсутствие или наличие заболевания у пациентов в медицинских исследованиях; наличие собственности, вкладов в банке при построении скоринговых систем для оценки возможности кредитования или страхования; ответы да/нет при обработке результатов анкетирования в социологических

опросах и пр. [2].

Рассчитанные в результате логистического анализа величины переменных в большинстве случаев принимают не точные значения 0 или 1, а значения в интервале. В этом случае полученные в ходе анализа результаты интерпретируют не как конкретные значения, а как вероятность P того, что результат может быть отнесен к определенному классу. Например, вероятность близкая к единице, позволит сделать вывод, что пациент находится в зоне риска (вероятнее болен, чем здоров) и для уточнения диагноза необходимы клинические исследования [3].

На рис.1 представлена часть исходных данных, на основании которых строилась модель.

	A	B	C	D	E	F	G
1	Сладж	Травмы	Физическая активность	Посты, голодание	ИМТ	Наследственность	Возраст
2	да	да	да	да	2	да	1
3	да	нет	да	да	1	да	2
4	нет	нет	да	нет	0	нет	1
5	нет	нет	да	нет	1	нет	1
6	нет	нет	да	нет	0	да	1
7	да	да	нет	да	0	да	3
8	нет	нет	нет	нет	0	нет	4
9	да	да	нет	да	2	нет	4
10	да	да	нет	да	3	нет	4
11	да	да	да	да	2	нет	3
12	да	да	нет	нет	3	нет	5
13	да	да	нет	да	1	да	3
14	нет	нет	да	да	1	нет	5
15	нет	нет	нет	нет	0	нет	1
16	да	нет	нет	да	1	нет	3
17	да	нет	нет	да	1	нет	1
18	да	да	нет	нет	1	да	4
19	да	нет	нет	да	1	да	3
20	нет	нет	да	нет	0	нет	5
21	да	да	нет	нет	1	нет	2
22	да	нет	нет	да	1	нет	3
23	нет	нет	нет	да	1	нет	3
24	да	нет	нет	нет	0	нет	3
25	да	нет	нет	нет	1	да	3
26	да	да	да	нет	1	да	3
27	нет	да	да	нет	0	нет	5
28	да	да	нет	нет	1	нет	2
29	да	да	нет	нет	1	да	3
30	да	да	нет	да	1	да	4
31	да	да	да	да	0	нет	2
32	нет	да	да	да	0	нет	2
33	да	да	нет	да	0	нет	5

Рисунок 1 – Исходные данные модели

Показанные на рис.1 данные первоначально были размещены в Microsoft Excel. В столбце **A** указан установленный диагноз (да – подтверждено наличие заболевания, 0 – нет); в столбце **B** – наличие ранее травм; в столбцах **C** и **D** – данные по физической активности и резким изменениям в режиме питания. ИМТ был классифицирован следующим образом : 0 – нормальное телосложение или дефицит массы; 1 – первая стадия ожирения, 2 и 3 – соответственно вторые и третьи стадии. ИМТ определяли как отношение веса в килограммах к росту в метрах.

В норме ИМТ равен 20÷25 кг/м^2; при ожирении I степени – 25÷30 кг/м^2; при ожирении II степени – 30÷40 кг/м^2; при ожирении III степени – более 40 кг/м^2. При дефиците массы тела ИМТ равен 18÷20 кг/м^2.

В столбце **F** – данные о наследственной предрасположенности к заболеванию.

Первичная обработка статистических данных позволила произвести градацию обследуемых по возрастному признаку (столбец **G** таблицы) .

Возраст до 25 лет обозначали как 1 (первая группа); от 25 до 34 лет – вторая группа; 3-ая группа – от 35 до 44 лет, 4-ая – от 45 до 54 лет; 5-ая группа – старше 54 лет.

Для обработки использовали пакет Statistica, модуль Nonlinear Estimate, в котором выбрана категория - **Logistic Regression**.

Для данных, представленных на рис.1 в качестве ***y*** рассматривали наличие заболевания (***y***=1 соответствует состоянию «сладж=да» и 0 в противном случае), в качестве независимых переменных значения в столбцах ($\mathbf{x_1}$ характеризует наличие травм у пациента в анамнезе; $\mathbf{x_2}$ наличие или отсутствие физической активности; $\mathbf{x_3}$ – зафиксированные изменения в режимах питания; $\mathbf{x_4}$ – ИМТ; $\mathbf{x_5}$ **-** наличие наследственной предрасположенности к заболеванию; $\mathbf{x_6}$ **-** возрастная категория).

При построении логистической модели определяются параметры уравнения вида

$$y = \frac{e^{b_0+b_1\cdot x_1+b_2\cdot x_2+\ldots b_n\cdot x_n}}{1+e^{b_0+b_1\cdot x_1+b_2\cdot x_2+\ldots b_n\cdot x_n}} \qquad (1)$$

где b_0, $b_1 \ldots b_n$ - коэффициенты логистической регрессии, n - количество независимых переменных.

Исходя из формулы (1) можно определить вероятность наступления некоторого события, в нашем случае это вероятность подтверждения диагноза, по формуле:

$$P = \frac{1}{1+e^{-(b_0+b_1\cdot x_1+b_2\cdot x_2+\ldots b_n\cdot x_n)}}$$

Для решения задач логистической регрессии в пакете Statistica используется метод максимального правдоподобия.

В результате расчетов получены параметры модели:

Таблица 1. Коэффициенты логистической регрессии

b_0	b_1	b_2	b_3	b_4	b_5	b_6
-0.80	17.98	-18.29	1.33	1.51	17.92	-0.17

На рис. 2 представлено результирующее окно расчетов параметров. Кроме значений коэффициентов логистической регрессии, которые отображаются в строке **Estimate**, в окне отображения результатов указан p-уровень гипотезы. Если p-уровень менее 5%, модель считается значимой. В нашем случае p=0.0001472, что позволило сделать вывод о

значимости разработанной модели.

Оценивая величину хи-квадрата Пирсона (Chi-square =26.97191 на рис. 2) и сравнивая полученное значение критерия хи-квадрат с критическим, получаем 26.97191 > 9.488, откуда делаем вывод о наличии статистически значимой зависимости параметра **У** от выбранных независимых переменных **X**. Уровень значимости данной взаимосвязи соответствует p<0.05.

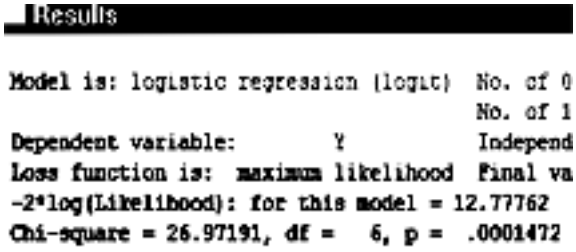
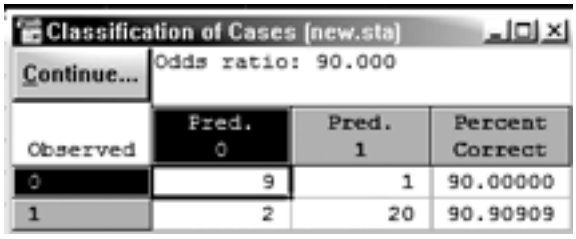


Рисунок 2 – Параметры модели логистической регрессии

Еще одним важным шагом в проверке качества построенной модели является оценка параметра *Отношение несогласия*. На рис. 3 показана таблица с числом наблюдений, которые были правильно или неправильно классифицированы в соответствии с полученной моделью [4]. В данной статье рассматривается подвыборка, состоящая из 22 человек, которые классифицированы как больные, и десяти человек – условно здоровые.



Observed	Pred. 0	Pred. 1	Percent Correct
0	9	1	90.00000
1	2	20	90.90909

Рисунок 3 – Принадлежность наблюдений к классам

Все наблюдения с предсказанными значениями (вероятностью) меньше или равными 0.5 классифицируются в пакете Statistica как неудача – «Failure», остальные, с предсказываемыми значениями больше 0.5, классифицируются как успех – «Success». *Отношение несогласия* вычисляется как отношение произведения чисел правильно классифицированных наблюдений к произведению чисел неправильно классифицированных результатов [5].

В столбце *Percent Correct* приведены проценты совпадений исходных и прогнозируемых данных. Так, в строке, соответствующей *Observed*=0 (по нашей классификации сладж не обнаружен), рассматривается процент правильной классификации здоровых пациентов. Следовательно, в 90.9 случаях из 100 здоровый пациент идентифицируется именно как здоровый, и менее, чем в 10 случаях из 100 он будет считаться больным ошибочно. Это объясняется тем, что даже у здорового человека существует определенный процент риска, обусловленный его возрастом, полом, массой тела, наследственной предрасположенностью, наличием серьезных травм, но далеко не все люди страдают этим заболеванием.

Во второй строке, соответствующей *Observed*=1 (подтверждено наличие сладжа), выводится процент правильно идентифицированных больных. В 90 случаях из 100 модель позволяет предсказать наличие заболевания и в 10 случаях из ста больной пациент будет считаться здоровым (значение Percent Correct=90), т.е болезнь не будет обнаружена на основании расчета вероятности риска по скоринговой модели.

Проведенные исследования и сопоставление полученных статистических данных и прогноза позволили считать возможным использование разработанной модели для начальной, доклинической диагностики.

Подставляя полученные коэффициенты логистической регрессии (см. табл. 1) в уравнение (1) можно получить прогнозируемые значения величины $Y_{теор}$.

Отметим, что если в исходном наборе данных зависимая переменная Y принимает значения 0 или 1 (да/нет), то результатом расчетов по полученной зависимости является массив данных в интервале от 0 до 1.

Например, для приводимого в табл.1. набора данных («нет» кодировалось как ноль, «да» – как единица), теоретические значения Y имели вид, показанный в табл. 2. В таблице приводятся первые десять данных выборки.

В табл. 2 в строке $Y_{теор}$ показана предсказанная вероятность того, что пациента можно отнести к классу больных, через Y обозначены реальные данные как в виде варианта да/нет, так и их бинарные значения.

Как видим, далеко не все значения во второй строке таблицы равны нулю или единице. В большинстве случаев значения колеблются в указанном интервале. В связи с этим возникает вопрос, следует ли считать больным, например, пациента под номером 5 с рассчитанной вероятностью 0.208 или пациента под номером 2 ($Y_{теор}$ =0.792).

Таким образом, при расчете значения $Y_{теор}$ по уравнению логистической регрессии (1), возникает проблема, каким образом должны

быть преобразованы полученных числовые значения $Y_{теор}$ в бинарные классификаторы.

Таблица 2. Коэффициенты логистической регрессии

№	Сладж (Y)		$Y_{теор}$
1	да	1	1
2	да	1	0.79173023
3	нет	0	0.0000000043
4	нет	0	0.0000000196
5	нет	0	0.20829880
6	да	1	1
7	нет	0	0.18388534
8	да	1	1
9	да	1	1
10	да	1	0.93882555

Следовательно, необходимо определить такое пороговое значение P, при котором выполняется система неравенств:

$$\begin{aligned} y \geq P_{порог} &\rightarrow пациент\ \ болен \\ y \leq P_{порог} &\rightarrow пациент\ \ здоров \end{aligned} \qquad (2)$$

Т.е. возникает вопрос, как определить пороговое значение вероятности, которое разделяет исследуемое множество на два класса: здоров/болен; кредитоспособен/ не кредитоспособен, эффективно лекарство для группы пациентов или нет. Причем от величины принятого порогового значения зависит точность модели и результаты предварительного диагностирования пациента.

Пороговое значение изменяется в интервале от нуля до единицы. Варьируя значение порогового P, каждый раз при расчетах будут получаться новые значения бинарного классификатора. В нашем случае, это то значение вероятности $P_{порог}$, которое следует использовать при разделении пациентов на больных и здоровых, формула (2).

Например, если в качестве порогового значения выбрать величину 0.5, то возникает проблема частичного отсечения больных пациентов, набравших недостаточно баллов риска.

Для решения этой проблемы использовали аппарат ROC (**R**eceiver **O**perating **C**haracteristic) – анализа.

ROC – анализ долгое время использовали в теории обработки сигналов и радиолокации для описания события правильного обнаружения сигнала и исключения ложных срабатываний. В настоящее время ROC – анализ нашел широкое применение не только в задачах медицинской диагностики, но и в банковском, кредитном скорринге, теории принятия решений.

ROC-кривая показывает зависимость количества верно классифицированных наблюдений со значением параметра, равным нулю, от количества неверно классифицированных наблюдений со значением параметра, равным единице. При этом предполагается, что у классификатора имеется параметр, варьируя который, будет происходить разделение на два класса. Этот параметр часто называют порогом, или точкой отсечения (cut-off value). В зависимости от него будут получаться различные значения *ошибок I и II рода*.

На основании данных, приведенных на рис. 3, можно определить значения, необходимые для проведения ROC – анализа, а именно:

TP (***T**rue **P**ositives*) – верно классифицированные случаи с заболеванием, для рассматриваемого примера TP=20;

TN (***T**rue **N**egatives*) – верно классифицированные случаи с отсутствием заболевания, для рассматриваемого примера TN=9;

FN (***F**alse **N**egatives*) – случаи с заболеванием, но при прогнозировании оно выявлено не было (*ошибка I рода*), так называемый «ложный пропуск» – болезнь не была обнаружена, для рассматриваемого примера FN=2;

FP (***F**alse **P**ositives*) – случаи, когда было вынесено решение о наличии заболевания, хотя его не было (*ошибка II рода*). Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его наличии, для рассматриваемого примера FP=1.

Для построения ROC-кривых необходимо вычислить два параметра: чувствительность (**Se**nsitivity) и специфичность (**Sp**ecificity).

Чувствительность определяли по формуле:

$$SE = \frac{TP}{TP + FN} \cdot 100\% \qquad (3)$$

Специфичность считали как:

$$SP = \frac{TN}{TN + FP} \cdot 100\% \qquad (4)$$

Эти два параметра используются для оценки требуемого уровня точности решения поставленной задачи. С медицинской точки зрения существуют следующие правила:

– если ставится задача максимального предотвращения пропуска больных, то должен

быть выбран высокий уровень чувствительности SE (гипердиагностика);

– если же лечение связано с серьезными побочными эффектами, то гипердиагностика не допустима и следует использовать более высокие уровни специфичности SP [6].

В первом случае увеличивается риск того, что здоровым пациентам будет поставлен диагноз «болен», во втором - риск не обнаружить болезнь

В табл.3. показаны рассчитанные значения перечисленных выше параметров TP, TN, FN, FP и величин чувствительности и специфичности, которые будут использованы при построении ROC – кривой. Параметры определялись для каждого значения порога отсечения $P_{порог}$ в пределах от 0 до 1 с выбранным шагом 0.1.

Кроме этого определяли величину FPR – процент ложно обнаруженных случаев заболевания, где

$$FPR = \frac{FP}{TN + FP} \cdot 100\% . \qquad (5)$$

При построении ROC-кривой по оси Y откладывают чувствительность SE, по оси X – значения параметра FPR.

Таблица 3. Параметры для построения ROC - кривой

№	$P_{порог}$	TP	TN	FN	FP	SE	SP	FPR
1	0.1	22	3	0	7	100%	30%	70%
2	0.2	22	4	0	6	100%	40%	60%
3	0.3	21	6	1	4	95%	60%	40%
4	0.4	20	8	2	2	91%	80%	20%
5	**0.5**	**19**	**9**	**3**	**1**	**86%**	**90%**	**10%**
6	**0.6**	**19**	**9**	**3**	**1**	**86%**	**90%**	**10%**
7	**0.7**	**19**	**9**	**3**	**1**	**86%**	**90%**	**10%**
8	**0.8**	**19**	**9**	**3**	**1**	**86%**	**90%**	**10%**
9	**0.9**	**17**	**9**	**5**	**1**	**77%**	**90%**	**10%**
10	1	0	10	22	0	0%	100%	0%

На рис. 4 показана построенная ROC – кривая.

Идеальная ROC-кривая должна проходить через верхний левый угол, в котором процент правильно идентифицированных случаев составляет 100% (идеальная чувствительность), а процент ошибки равен нулю.

Ясно, что в реальности построенная кривая должна стремиться к идеальной, однако, вряд ли будет достигать ее.

Отсюда следует, что чем ближе ROC - кривая к верхнему левому углу, тем выше уровень качества прогнозирования разработанной модели. И наоборот, чем меньше изгиб кривой и чем ближе она расположена к диагональной прямой (прямая линия под углом 45°, соединяющая левый нижний и правый верхний углы, на рис.4 показана пунктиром), тем менее эффективна модель.

Близость ROC-кривой к диагональной линии соответствует так называемому «бесполезному» классификатору, т.е. полной неразличимости классов [7,8]. А, следовательно, свидетельствует о не соответствии разработанной модели полученным статистическим данным.

Как можно заметить, построенная по разработанной модели ROC – кривая достаточно близка к идеальной.

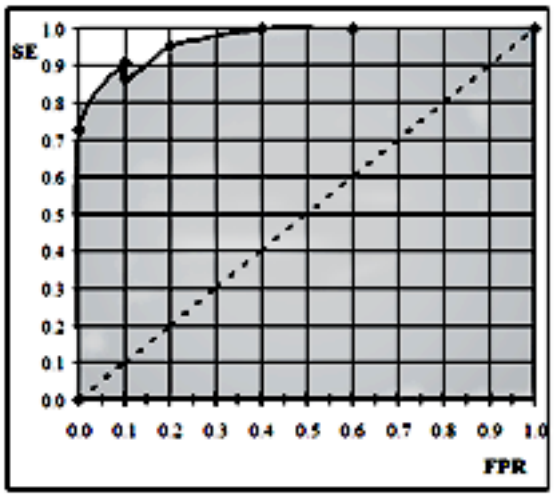


Рисунок 4 – ROC - кривая

Визуальная оценка кривых ROC не всегда позволяет выявить наиболее целесообразное значение порога и численно оценить качество разработанной модели.

Часто для оценки качества ROC-кривых используют оценку площади под кривыми (на рис. 4 закрашена серым цветом). Эта оценка может быть получена непосредственно вычислением площади под многогранником, ограниченным справа и снизу осями координат и слева вверху – экспериментально полученными точками (рис. 4).

Численный показатель площади под кривой называется AUC (**A**rea **U**nder **C**urve).

В литературе приводится экспертная шкала для значений AUC, по которой можно судить о качестве разработанной модели [9,10], см. табл. 4.

Нами было вычислено значение AUC, величина составила более 96%, что позволяет сделать вывод об «отличном» качестве модели.

Отметим, что на первом этапе

исследований нами не рассматривались такие факторы, как наследственность и возраст. Полученные результаты (площадь около 83%) говорили об «Очень хорошем» качестве. Введение дополнительных параметров позволило улучшить качество модели.

Таблица 4. Интервалы площади под ROC-кривой и оценка качества модели

Интервал 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	Неудовлетворительное

Наряду с вышеперечисленными факторами, на первом этапе исследований рассматривали влияние пола, приема определенного вида лекарств на вероятность возникновения болезни. Однако, эти факторы не оказали существенного влияния на точность модели и были исключены из рассмотрения.

Как видно из данных табл. 3 и рис. 4, значения рассматриваемых параметров SE, SP, FPR не меняются для порога, находящегося в пределах

$$0.5 \le P_{порог} \le 0.9 \quad . \qquad (6)$$

Соответствующие строки выделены в табл. 3. Это означает, что пороговые значения могут быть любым числом в указанном интервале.

Неоднозначность полученного ответа обусловила необходимость в дополнительных исследованиях с целью уточнения порогового значения.

В качестве дополнительной проверки качества модели и уточнения интервала порогового значения определяли отношение суммы правильно идентифицированных случаев (как при отсутствии болезни, так и в случае ее наличия) к общему числу имеющихся экспериментальных данных. Данный параметр определили как переменную F= (TP+TN)/N.

Результаты зависимости переменной F от величины порога показаны в табл. 5.

Для уточнения значения порогового значения была поставлена и решена задача подбора параметров аппроксимирующей функции, описывающей зависимость числа правильно идентифицированных случаев от величины порога.

Для решения задачи использовали модуль Nonlinear Estimate пакета Statistica. Проведен подбор коэффициентов полиномиального уравнения. Как показала оценка погрешности подбора, полином пятой степени описывает данные с наименьшей погрешностью [11].

Таблица 5. Зависимость параметра F от величины порога

$P_{порог}$	F
0.1	81%
0.2	88%
0.3	91%
0.4	91%
0.5	91%
0.6	91%
0.7	91%
0.8	88%
0.9	81%

На рис. 5 показано результирующее окно пакета Statistica .

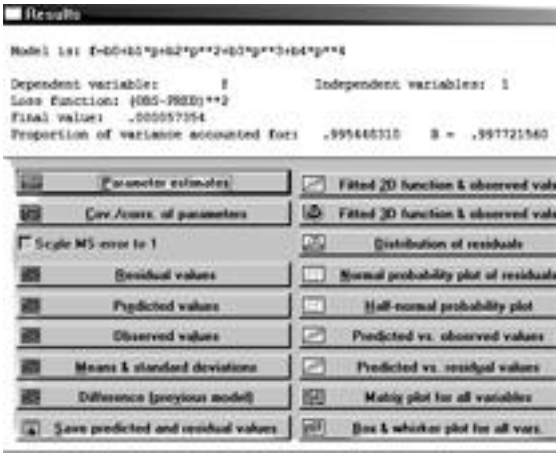


Рисунок 5 – Расчет параметров полиномиальной функции в пакете Statistica

Зависимость переменной F от величины порога можно описать уравнением вида:

$$F=0.67+1.90\times P_{порог}-5.72\times P_{порог}^{2}+7.64\times P_{порог}^{3}-3.82 P_{порог}^{4} \qquad (7)$$

Дифференцируя функцию F, находим точку экстремума. Как показали расчеты, при значении порога $P_{порог}=0.62$ имеет место максимальное количество правильно диагностированных случаев, равное 91%. Это значение является уточненным значением порога, см. формулу (6).

Таким образом, алгоритм идентификации пациентов свелся к нескольким шагам:

1. ввести значения физических характеристик обследуемого (возраст, значение ИМТ и пр. факторы, указанные в табл.1);
2. рассчитать по уравнению

логистической регрессии значения Yteop $\frac{e^{-0.8+17.98\cdot x_1-18.29\cdot x_2+1.33\cdot x_3+1.54\cdot x_4+17.92\cdot x_5-0.17\cdot x_6}}{1+e^{-0.8+17.98\cdot x_1-18.29\cdot x_2+1.33\cdot x_3+1.54\cdot x_4+17.92\cdot x_5-0.17\cdot x_6}}$;

3. сравнить полученные значения с пороговым значением (в нашем примере 0.62);

4. если Yteop > 0.62 можно сделать вывод о высокой вероятности заболевания и необходимости клинического диагностирования;

5. в противном случае следует считать, что обследуемый скорее здоров, чем болен.

Выводы

В работе рассмотрены вопросы разработки скоринговой модели, позволяющей по предварительному анкетированию пациентов без дополнительной клинической диагностики оценивать степень возможного риска заболевания.

Выявлены факторы, оказывающее влияние на вероятность возникновения заболевания. Проведена оценка качества модели. Точность проведенных расчетов более 91%, что позволяет рекомендовать предложенную методику для составления скоринг - карт пациентов.

Литература

1. Мильчаков К.С., Шебалков М.П. Скоринговые карты в медицине: обзор и анализ публикаций. // Врач и информационные технологии, №1, 2015. - с.71-79.

2. Методы статистической обработки медицинских данных: Методические рекомендации для ординаторов и аспирантов медицинских учебных заведений, научных работников. / сост.: Кочетов А.Г., Лянг О.В., Масенко В.П., Жиров И.В., Наконечников С.Н., Терещенко С.Н. – М.: РКНПК, 2012. – 42 с.

3. Леонов В. Логистическая регрессия в медицине и биологии / Биометрика. Журнал для медиков и биологов, сторонников доказательной биомедицины. - Режим доступа: http://www.biometrica.tomsk.ru/logit_1.htm.

4. Бэстенс Д.-Э., Ван Ден Берг В.-М., Вуд Д. Нейронные сети и финансовые рынки. Принятие решений в торговых операциях. - М.: Научное издательство ТВП, 1997. - 235с.

5. Логистическая регрессия / Портал Знаний StatSoft. - Режим доступа: http://statistica.ru/theory/logisticheskaya-regressiya/

6. Логистическая регрессия и ROC-анализ - математический аппарат / Base Group Lab. - Режим доступа: https://basegroup.ru/community/articles/logistic.

7. Бююль А., Цёфель П. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: - Пер. с нем. - СПб.: ДиаСофтЮП, 2005. - 608 с.

8. Логистическая регрессия. / Центр статистического анализа. - Режим доступа: http://statmethods.ru/konsalting/statistics-metody/116-logisticheskaya-regressiya.html .

9. Цыплаков А.А. Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии. Методическое пособие. - Новосибирск: НГУ, 1997. - 129 с.

10. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях. - М.: ГЕОТАРDМЕД, 2001. - 256 с.

11. Боровиков В.П., Боровиков И.П. Statistica. Статистический анализ и обработка данных в среде Windows. -М.: Информационно-издательский дом "Филинъ", 1997.- 608с.

Анохина И.Ю. Разработка скоринговой модели с использованием методов логистической регрессии и ROC – анализа. *Рассмотрены вопросы создания скоринговой модели на основании данных медицинских исследований. При моделировании использованы методы интеллектуального анализа данных, к которым относится аппарат логистического регрессионного анализа, логико-статистические подходы к распознаванию образов, в том числе ROC – анализ. Описаны этапы разработки скоринговой модели. Проведена оценка точности моделирования.*

Ключевые слова: *статистические данные, скоринговые модели, методология разработки, риски, логистическая регрессия, ROC – анализ, пакет Statistica.*

Anokhina I.Y. *Development of a scoring model using logistic regression and ROC methods - analysis. The problems of creating a scoring model based on medical research. The simulation used the methods of data mining, which includes the device logistic regression analysis, logical-statistical approaches to pattern recognition, including ROC - analysis. We describe the stages of development of the scoring model. The accuracy of the simulation.*

Keywords: *statistics, scoring models, methodology development, risks, logistic regression, ROC - analysis, Statistica package.*

Статья поступила в редакцию 20.11.2016
Рекомендована к публикации д-ром техн. наук В.Н. Павлышом