

Министерство науки и высшего образования РФ
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Пензенский государственный университет»

Факультет вычислительной техники
Кафедра математики и суперкомпьютерного моделирования

ОТЧЕТ

по производственной (преддипломной) практике
«Применение метода ROC-анализа для оценки качества модели
логистической регрессии»

Выполнил: студент гр. 19ВМ1

Симакова Е.А.

Руководитель: Мартынов И.В.

Пенза 2023

Министерство науки и высшего образования РФ
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Пензенский государственный университет»

«УТВЕРЖДАЮ»

Зав. кафедрой МСМ, д.ф.-м.н.

_____ Ю.Г. Смирнов

ЗАДАНИЕ

на производственную (преддипломную) практику

(8 семестр)

Вариант 4

Тема «Применение метода ROC-анализа для оценки качества модели
логистической регрессии»

1. Изучение теоретического материала по заданной теме;
2. Поиск исходных данных для проведения исследования;
3. Получение и анализ численных результатов.

Руководитель работы

Мартынов И.В.

Содержание

Введение	4
Глава I. Теоретическая часть	6
§1. Логистическая регрессия.....	6
§2. ROC-анализ.....	8
Глава II. Практическая часть	12
§1. Диагностика сахарного диабета	12
Заключение.....	23
Список литературы	24

Введение

Практическая работа посвящена применению метода ROC-анализа для оценки качества модели логистической регрессии в задачах диагностики сахарного диабета и выявления неблагонадежных потенциальных заемщиков.

Наряду с дисперсионным и регрессионным анализами, логистическая регрессия занимает значительное место, как разновидность множественной линейной регрессии. Ее общее назначение состоит в анализе связи между несколькими независимыми переменными (предикторами) и зависимой переменной (откликом). Логистическая регрессия применяется в том случае, когда зависимая переменная является бинарной (т.е. она может принимать только два значения). Прогнозируемым откликом в этом случае будет являться вероятность наступления положительного события (исхода) для конкретного испытуемого. Возможным вариантом решения является непрерывная величиной, значение которой находится в диапазоне от 0 до 1. Значение зависимой переменной, близкое к 0, указывает на малую вероятность наступления интересующего нас исхода, а значение, близкое к 1, указывает, наоборот, на высокую вероятность наступления этого исхода [1].

В качестве инструмента оценки диагностической способности модели логистической регрессии используется ROC-анализ. Он широко применяется в таких областях, как: маркетинг, кредитный скоринг, военное дело, биомедицина и в других областях, где присутствует бинарная вероятностная классификация. В задачах бинарной классификации, когда модель предсказывает вероятность того, что наблюдения принадлежат к одному из двух классов, очень важно выбрать точку отсечения, то есть порог вероятности, разделяющий эти два класса.

Выбирая точку отсечения, мы контролируем вероятность правильного распознавания положительных и отрицательных примеров. По мере уменьшения порога отсечения вероятность ошибочного распознавания положительных наблюдений возрастает, а по мере его увеличения возрастает и вероятность неправильного распознавания отрицательных наблюдений.

Цель заключается в том, чтобы подобрать такое значение точки отсечения, которое дает наибольшую точность распознавания заданного класса, а какого именно – определяется постановкой задачи. Для этого и используется ROC-анализ [2].

Глава I. Теоретическая часть

§1. Логистическая регрессия

Логистическая регрессия отличается от множественной тем, что в качестве зависимой переменной используется не количественная, а бинарная переменная, которая может принимать лишь два возможных значения. В большинстве случаев эти два значения обозначают принадлежность или не принадлежность объекта к какой-либо группе.

Если множественная регрессия позволяет предсказывать количественное значение зависимой переменной (критерия) основываясь на известных значениях независимых переменных (предикторов), то логистическая регрессия прогнозирует вероятность исхода некоторого события, находящуюся в пределах от 0 до 1.

С логистической регрессией связаны следующие математические понятия: вероятность, шанс и натуральный логарифм шанса (логит). Вероятность – это ожидаемая относительная частота некоторого события. Шанс – это отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет [3].

Ключевым параметром логистической регрессии является логит. Логит равен натуральному логарифму шанса.

Уравнение множественной регрессии, имеет следующий вид:

$$Z = B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n. \quad (1.1)$$

Уравнение логистической регрессии похоже на уравнение множественной регрессии:

$$\ln \left[\frac{P}{1-P} \right] = B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n. \quad (1.2)$$

В этом уравнении P – вероятность наступления положительного исхода, $\frac{P}{1-P}$ – шанс наступления положительного исхода, $\ln \left[\frac{P}{1-P} \right]$ – логит.

Более простая запись уравнения логистической регрессии выглядит следующим образом:

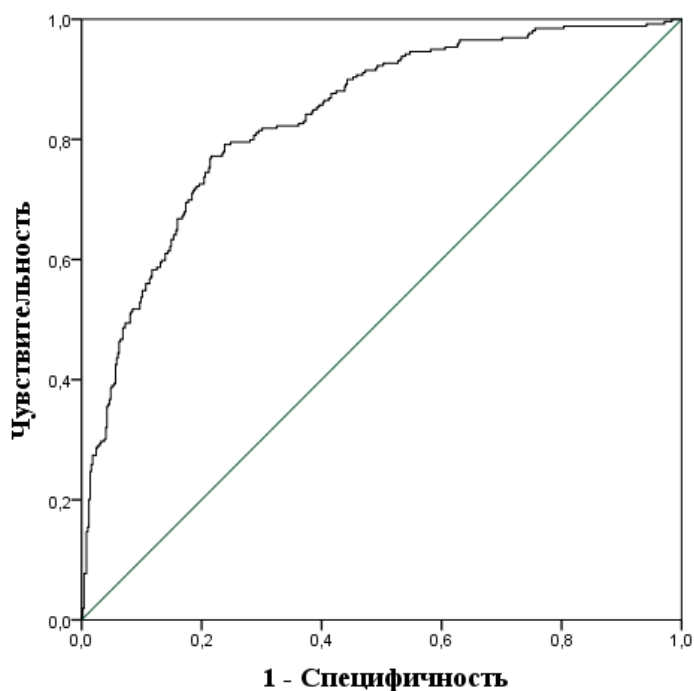
$$P = \frac{1}{1 + e^{-z}}. \quad (1.3)$$

§2. ROC-анализ

ROC-анализ тесно связан с логистической регрессией, являясь заключительным этапом при разработке наиболее адекватной регрессионной модели. Применяется же ROC-анализ для оценки качества созданной модели. Он позволяет проанализировать два основных показателя: чувствительность и специфичность, подобрать порог (точку) отсечения и выбрать регрессионную модель с наилучшей прогностической силой.

Так как результат решения задачи по логистической регрессионной модели не дает однозначного ответа, а лишь предлагает вероятность наступления одного из двух заданных состояний, то всегда важно знать, какое именно решение необходимо принять в той или иной ситуации. В связи с этим было принято, что прогноз положительного эффекта дается при вероятности $> 0,5$, а отрицательного – при вероятности $\leq 0,5$.

Для того, чтобы получить наиболее точное решение поставленной задачи, которое соответствует конкретной ситуации, используется ROC-кривая. ROC-кривая – это график, позволяющий дать оценку качеству модели по разделению ее на два класса [1]:



Так как классов два, то один из них называется классом с положительными исходами, а второй – с отрицательными исходами. ROC-

кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. При этом предполагается, что у классификатора есть некоторый параметр, изменяя который, можно получить то или иное разбиение на два класса. Этот параметр называют порогом, или точкой отсечения, которая наиболее адекватно отсекает одну диагностируемую группу от другой. Принято, что этой точкой является точка, которая наиболее близко расположена к верхнему левому углу графика или максимально удалена от диагональной прямой линии.

Кроме графика ROC-кривой, для оценки качества регрессионной модели используется такая характеристика, как площадь под кривой ROC AUC. Можно считать, что чем больше показатель AUC, тем лучшей предсказательной способностью обладает модель.

Для общего понимания интерпретации оценок качества тестируемого метода диагностики приняты определенные характеристики. Исходные данные, на основе которых рассчитываются характеристики метода, задаются в виде таблицы:

Решение по тестируемому методу	Фактическое состояние объектов	
	Положительное	Отрицательное
Положительное	TP	FP
Отрицательное	FN	TN

Определения характеристик следующие:

- TP (*True Positives*) – верно классифицированные положительные примеры (истинно положительные случаи);
- TN (*True Negatives*) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);
- FN (*False Negatives*) – положительные примеры, которые классифицируются как отрицательные (ошибка I рода). Это своего рода

«ложный пропуск» – когда интересующее нас событие ошибочно не обнаруживается (ложноотрицательные случаи);

- FP (*False Positives*) – отрицательные примеры, которые классифицируются как положительные (ошибка II рода). Это ложное обнаружение, когда интересующее нас событие ошибочно обнаруживается (ложноположительные случаи).

При изучении качества диагностического теста чаще всего пользуются не абсолютными, а относительными показателями – долями, выраженными в процентах, которые переходят в такие определения, как: чувствительность и специфичность метода.

Чувствительность – это доля истинно положительных случаев, которые были правильно определены тестируемым методом:

$$Se = \frac{TP}{TP + FN} \cdot 100\%. \quad (2.1)$$

Специфичность – это доля истинно отрицательных случаев, которые были правильно определены тестируемым методом:

$$Sp = \frac{TN}{TN + FP} \cdot 100\%. \quad (2.2)$$

Регрессионная модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные случаи). И наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные случаи).

Чтобы минимизировать ошибки I рода (ложноотрицательные случаи), нужно использовать регрессионную модель с высокой чувствительностью. Чтобы минимизировать ошибки II рода (ложноположительные случаи), нужно использовать регрессионную модель с высокой специфичностью.

Идеальная модель обладает 100% чувствительностью и специфичностью. Но на практике этого достичь невозможно, более того, невозможно повысить чувствительность и специфичность модели

одновременно. Компромисс находится с помощью порога отсечения, так как пороговое значение определяется соотношением чувствительности и специфичности. В таких случаях обычно принято говорить о задаче нахождения оптимального порога отсечения. Точка отсечения нужна для того, чтобы применять регрессионную модель на практике, а именно соотносить новые данные к одному из двух классов [1].

Для того, чтобы уточнить оптимальный порог нужно задать критерий для его определения, так как в различных задачах присутствует своя стратегия.

Критериями выбора порога отсечения могут выступать:

1. Требование минимальной величины чувствительности модели. К примеру, нужно обеспечить чувствительность теста не менее 80%. В этом случае оптимальной точкой отсечения будет максимальная специфичность при 80% чувствительности.
2. Требование максимальной суммарной чувствительности и специфичности модели:

$$Max_s = \max(Se + Sp). \quad (2.3)$$

3. Требование баланса между чувствительностью и специфичностью модели:

$$Min_d = \min|Se - Sp|. \quad (2.4)$$

Глава II. Практическая часть

§1. Диагностика сахарного диабета

Рассмотрим медицинскую задачу, заключающуюся в диагностике сахарного диабета. Скрининг на выявление диабета должен проводиться не только людям с наличием симптомов, но и пациентам без симптомов заболевания, чтобы диагностировать его на ранней стадии [4].

Для того, чтобы лучше понять исходные данные, дадим несколько определений:

Артериальное диастолическое давление (АД) – это давление, которое поддерживается в сосудах в момент расслабления (диастолы) сердца [5].

Измерение толщины кожной складки трехглавой мышцы (Толщина КС) проводится с целью оценки запасов жира в организме.

Индекс массы тела (ИМТ) – это величина, которая позволяет оценить степень соответствия массы человека и его роста и тем самым косвенно судить о том, является ли масса тела недостаточной, нормальной или избыточной [6].

Рассмотрим таблицу исходных данных, в которую вошли 754 наблюдения. Распределение зависимой переменной следующее: 495 случаев отсутствия заболевания, 259 – его наличие [7].

В таблице приняты следующие обозначения:

Предикторы:

X_1 – Оценка сбалансированности питания по двенадцатибалльной шкале (ОСП);

X_2 – Концентрация глюкозы в крови (Глюкоза);

X_3 – Артериальное диастолическое давление (АД);

X_4 – Толщина кожной складки трехглавой мышцы (Толщина КС);

X_5 – Двухчасовой сывороточный инсулин (Инсулин);

X_6 – Индекс массы тела (ИМТ);

X_7 – Числовой параметр наследственности диабета (Наследственность);

X_8 – Возраст.

Объясняемая переменная:

Y – Диагноз: 1 – наличие заболевания, 0 – отсутствие.

ОСП	Глюкоза	АД	Толщина КС	Инсулин	ИМТ	Наследственность	Возраст	Диагноз
6	148	72	35,0	0,0	33,6	0,627	50	1
1	85	66	29,0	0,0	26,6	0,351	31	0
8	183	64	0,0	0,0	23,3	0,672	32	1
1	89	66	23,0	94,0	28,1	0,167	21	0
0	137	40	35,0	168,0	43,1	2,288	33	1
5	116	74	0,0	0,0	25,6	0,201	30	0
3	78	50	32,0	88,0	31,0	0,248	26	1
10	115	0	0,0	0,0	35,3	0,134	29	0
2	197	70	45,0	543,0	30,5	0,158	53	1
8	125	96	0,0	0,0	0,0	0,232	54	1
4	110	92	0,0	0,0	37,6	0,191	30	0
10	168	74	0,0	0,0	38,0	0,537	34	1
10	139	80	0,0	0,0	27,1	1,441	57	0
1	189	60	23,0	846,0	30,1	0,398	59	1
5	166	72	19,0	175,0	25,8	0,587	51	1
7	100	0	0,0	0,0	30,0	0,484	32	1
0	118	84	47,0	230,0	45,8	0,551	31	1
7	107	74	0,0	0,0	29,6	0,254	31	1
1	103	30	38,0	83,0	43,3	0,183	33	0
1	115	70	30,0	96,0	34,6	0,529	32	1
3	126	88	41,0	235,0	39,3	0,704	27	0
8	99	84	0,0	0,0	35,4	0,388	50	0
7	196	90	0,0	0,0	39,8	0,451	41	1
9	119	80	35,0	0,0	29,0	0,263	29	1
11	143	94	33,0	146,0	36,6	0,254	51	1
10	125	70	26,0	115,0	31,1	0,205	41	1
7	147	76	0,0	0,0	39,4	0,257	43	1
1	97	66	15,0	140,0	23,2	0,487	22	0
5	117	92	0,0	0,0	34,1	0,337	38	0
5	109	75	26,0	0,0	36,0	0,546	60	0
3	158	76	36,0	245,0	31,6	0,851	28	1
3	88	58	11,0	54,0	24,8	0,267	22	0
6	92	92	0,0	0,0	19,9	0,188	28	0
10	122	78	31,0	0,0	27,6	0,512	45	0
4	103	60	33,0	192,0	24,0	0,966	33	0
11	138	76	0,0	0,0	33,2	0,420	35	0
9	102	76	37,0	0,0	32,9	0,665	46	1
2	90	68	42,0	0,0	38,2	0,503	27	1
4	111	72	47,0	207,0	37,1	1,390	56	1
3	180	64	25,0	70,0	34,0	0,271	26	0
7	133	84	0,0	0,0	40,2	0,696	37	0
7	106	92	18,0	0,0	22,7	0,235	48	0
9	171	110	24,0	240,0	45,4	0,721	54	1

7	159	64	0,0	0,0	27,4	0,294	40	0
0	180	66	39,0	0,0	42,0	1,893	25	1
1	146	56	0,0	0,0	29,7	0,564	29	0
2	71	70	27,0	0,0	28,0	0,586	22	0
7	103	66	32,0	0,0	39,1	0,344	31	1
7	105	0	0,0	0,0	0,0	0,305	24	0
1	103	80	11,0	82,0	19,4	0,491	22	0
1	101	50	15,0	36,0	24,2	0,526	26	0
5	88	66	21,0	23,0	24,4	0,342	30	0
8	176	90	34,0	300,0	33,7	0,467	58	1
7	150	66	42,0	342,0	34,7	0,718	42	0
1	73	50	10,0	0,0	23,0	0,248	21	0
7	187	68	39,0	304,0	37,7	0,254	41	1
0	100	88	60,0	110,0	46,8	0,962	31	0
0	146	82	0,0	0,0	40,5	1,781	44	0
0	105	64	41,0	142,0	41,5	0,173	22	0
2	84	0	0,0	0,0	0,0	0,304	21	0
8	133	72	0,0	0,0	32,9	0,270	39	1
5	44	62	0,0	0,0	25,0	0,587	36	0
2	141	58	34,0	128,0	25,4	0,699	24	0
7	114	66	0,0	0,0	32,8	0,258	42	1
5	99	74	27,0	0,0	29,0	0,203	32	0
0	109	88	30,0	0,0	32,5	0,855	38	1
2	109	92	0,0	0,0	42,7	0,845	54	0
1	95	66	13,0	38,0	19,6	0,334	25	0
4	146	85	27,0	100,0	28,9	0,189	27	0
2	100	66	20,0	90,0	32,9	0,867	28	1
5	139	64	35,0	140,0	28,6	0,411	26	0
4	129	86	20,0	270,0	35,1	0,231	23	0
1	79	75	30,0	0,0	32,0	0,396	22	0
1	0	48	20,0	0,0	24,7	0,140	22	0
7	62	78	0,0	0,0	32,6	0,391	41	0
5	95	72	33,0	0,0	37,7	0,370	27	0
0	131	0	0,0	0,0	43,2	0,270	26	1
2	112	66	22,0	0,0	25,0	0,307	24	0
3	113	44	13,0	0,0	22,4	0,140	22	0
2	74	0	0,0	0,0	0,0	0,102	22	0
7	83	78	26,0	71,0	29,3	0,767	36	0
0	101	65	28,0	0,0	24,6	0,237	22	0
5	137	108	0,0	0,0	48,8	0,227	37	1
2	110	74	29,0	125,0	32,4	0,698	27	0
2	100	68	25,0	71,0	38,5	0,324	26	0
1	107	68	19,0	0,0	26,5	0,165	24	0
1	80	55	0,0	0,0	19,1	0,258	21	0
4	123	80	15,0	176,0	32,0	0,443	34	0
7	81	78	40,0	48,0	46,7	0,261	42	0
4	134	72	0,0	0,0	23,8	0,277	60	1
...	...	...	...	...	...	...	...	...

Методом логистической регрессии будет прогнозироваться зависимая переменная Y на основе имеющихся предикторов.

Для построения модели логистической регрессии применим метод отношения правдоподобия, который заключается в том, что в регрессионное уравнение будут пошагово включаться предикторы, оказывающие наибольшее воздействие на зависимую переменную, до последнего предиктора, чье воздействие окажется значимым.

Шаг	Р-квадрат Нэйджелкерка
1	0,295
2	0,345
3	0,378
4	0,392
5	0,399

Из этой таблицы видно, что методу отношения правдоподобия понадобилось пять шагов для того, чтобы включить все нужные переменные в уравнение регрессии.

Р-квадрат Нэйджелкерка показывает долю влияния всех предикторов модели на дисперсию зависимой переменной. Значение Р-квадрата находится в промежутке от 0 до 1. Значение 0 указывает на то, что зависимая переменная вообще не может быть объяснена предикторной переменной. Значение 1 указывает на то, что зависимая переменная может быть полностью объяснена без ошибок с помощью переменной-предиктора [3].

Видно, что в модели с каждым шагом увеличивается доля объясненной дисперсии с 0,295 до 0,399.

Теперь оценим общее качество модели логистической регрессии с помощью таблицы классификации, в которой сравниваются прогнозируемые значения зависимой переменной и фактические наблюдаемые значения.

Наблюдаемые			Предсказанные		
			Диагноз		Процент правильных
			0	1	
Шаг 1	Диагноз	0	440	55	88,9
		1	133	126	48,6
	Общая процентная доля				75,1
Шаг 2	Диагноз	0	443	52	89,5
		1	124	135	52,1
	Общая процентная доля				76,7
Шаг 3	Диагноз	0	435	60	87,9
		1	114	145	56,0
	Общая процентная доля				76,9
Шаг 4	Диагноз	0	439	56	88,7
		1	116	143	55,2
	Общая процентная доля				77,2
Шаг 5	Диагноз	0	439	56	88,7
		1	113	146	56,4
	Общая процентная доля				77,6

Из таблицы можно сделать вывод о том, что из общего числа здоровых пациентов, равного 495, являются таковыми только 439 человек. Остальных 56 называют «ложноположительными», так как они были признаны больными, хотя являются здоровыми. Из общего числа больных пациентов, равного 259, являются таковыми только 146 человек. Оставшихся 113 называют «ложноотрицательными», так как они были признаны здоровыми, хотя являются больными.

Таким образом, процент правильных предсказаний по здоровым составляет 88,7%, а процент правильных предсказаний по больным составляет 56,4%. В общем, правильно были распознаны 585 человек из 754, что составляет 77,6%.

		В	Значимость	Exp (В)
Шаг 1	Глюкоза	0,038	0,000	1,039
	Константа	-5,380	0,000	0,005
Шаг 2	Глюкоза	0,035	0,000	1,036
	ИМТ	0,075	0,000	1,078
	Константа	-7,496	0,000	0,001
Шаг 3	ОСП	0,142	0,000	1,153
	Глюкоза	0,034	0,000	1,035
	ИМТ	0,082	0,000	1,085
	Константа	-8,154	0,000	0,000
Шаг 4	ОСП	0,147	0,000	1,158
	Глюкоза	0,034	0,000	1,035
	ИМТ	0,078	0,000	1,081
	Наследственность	0,926	0,002	2,525
	Константа	-8,462	0,000	0,000
Шаг 5	ОСП	0,160	0,000	1,173
	Глюкоза	0,035	0,000	1,035
	АД	-0,012	0,019	0,988
	ИМТ	0,085	0,000	1,089
	Наследственность	0,935	0,002	2,546
	Константа	-8,012	0,000	0,000

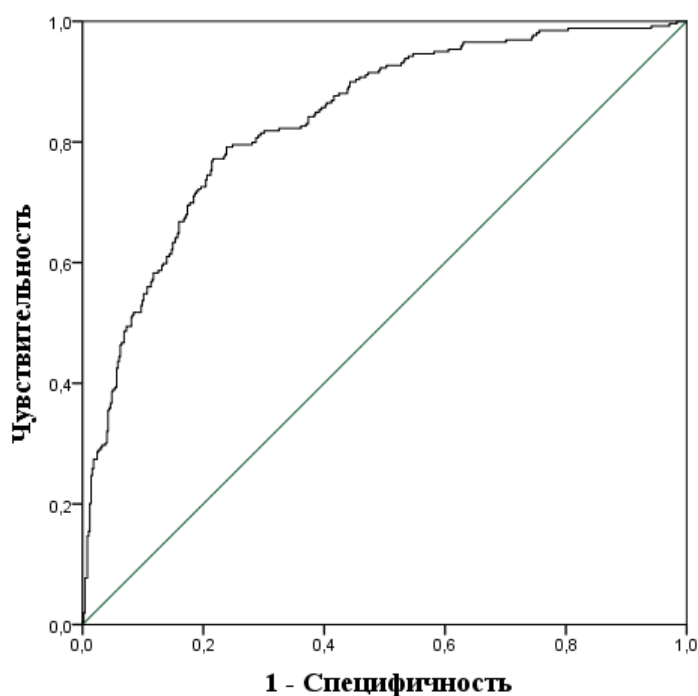
Данная таблица показывает эффекты включения переменных в уравнение логистической регрессии на каждом шаге его построения. Строка Константа для каждого шага соответствует константе B_0 регрессионного уравнения. В – это коэффициенты уравнения регрессии. Значимость показывает является ли коэффициент в модели статистически значимым или нет. Если значимость $< 0,05$, то коэффициент является статистически значимым, а если значимость $\geq 0,05$, то коэффициент статистически незначим. Exp (В) – это шанс, который равен отношению вероятности того, что событие произойдет, к вероятности того, что событие не произойдет [3].

Таким образом из всех предикторов в результате применения метода отношения правдоподобия было выбрано пять статистически важных

переменных: X_1 – ОСП, X_2 – Глюкоза, X_3 – АД, X_6 – ИМТ, X_7 – Наследственность.

Для того, чтобы улучшить общее качество модели логистической регрессии применим ROC-анализ.

Для того, чтобы получить наиболее точное решение поставленной задачи, которое соответствует конкретной ситуации, построим ROC-кривую:



Выберем точки отсечения для поставленной задачи:

1. Если мы выберем точку 0,138, в котором чувствительность диагностического теста будет очень высокой (95%), то получится гипердиагностика пациентов – предотвращение пропуска больных будет максимальным.

Наблюдаемые			Предсказанные		
			Диагноз		Процент правильных
			0	1	
Шаг 1	Диагноз	0	128	367	25,9
		1	11	248	95,8
	Общая процентная доля				49,9
Шаг 2	Диагноз	0	165	330	33,3
		1	10	249	96,1
	Общая процентная доля				54,9
Шаг 3	Диагноз	0	188	307	38,0

		1	13	246	95,0
	Общая процентная доля				57,6
Шаг 4	Диагноз	0	203	292	41,0
		1	12	247	95,4
	Общая процентная доля				59,7
Шаг 5	Диагноз	0	207	288	41,8
		1	13	246	95,0
	Общая процентная доля				60,1

Повторяя процедуру логистической регрессии, построим модель с новой точкой отсечения 0,138. Получится новая модель, которая обладает следующей предикторной способностью: чувствительность – 95%, специфичность – 41,8%, общее качество модели – 60,1%. Таким образом модель ухудшилась на 17,5%.

2. Выберем точку 0,328, в которой чувствительность и специфичность диагностического теста будут примерно совпадать (точка баланса).

Наблюдаемые			Предсказанные		
			Диагноз		Процент правильных
			0	1	
Шаг 1	Диагноз	0	356	139	71,9
		1	76	183	70,7
	Общая процентная доля				71,5
Шаг 2	Диагноз	0	363	132	73,3
		1	73	186	71,8
	Общая процентная доля				72,8
Шаг 3	Диагноз	0	370	125	74,7
		1	69	190	73,4
	Общая процентная доля				74,3
Шаг 4	Диагноз	0	380	115	76,8
		1	63	196	75,7
	Общая процентная доля				76,4
Шаг 5	Диагноз	0	382	113	77,2
		1	59	200	77,2
	Общая процентная доля				77,2

Повторяя процедуру логистической регрессии, построим модель с новой точкой отсечения 0,328. Получится новая модель, которая обладает

следующей предикторной способностью: чувствительность – 77,2%, специфичность – 77,2%, общее качество модели – 77,2%. Таким образом модель ухудшилась на 0,4%.

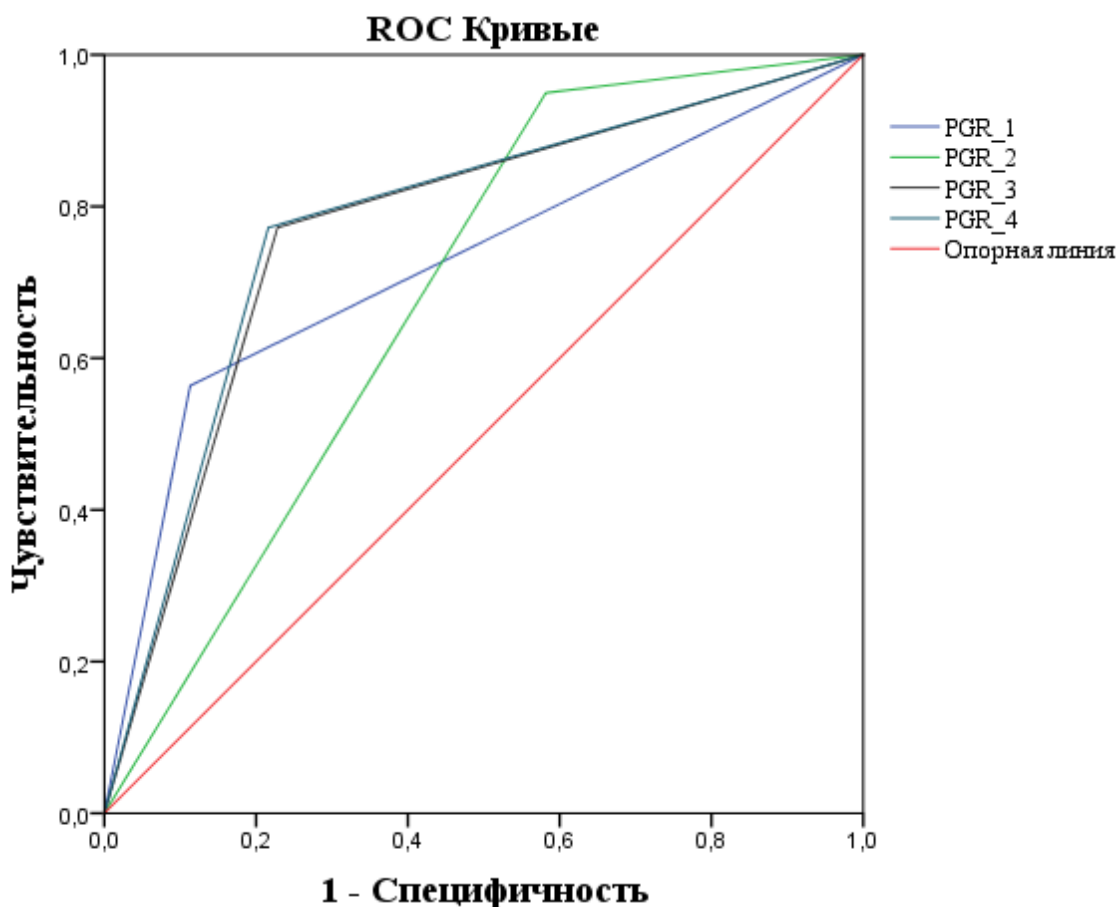
3. Выберем точку 0,332, которая обеспечивает максимум чувствительности и специфичности диагностического теста.

Наблюдаемые			Предсказанные		
			Диагноз		Процент правильных
			0	1	
Шаг 1	Диагноз	0	363	132	73,3
		1	78	181	69,9
	Общая процентная доля				72,1
Шаг 2	Диагноз	0	364	131	73,5
		1	75	184	71,0
	Общая процентная доля				72,7
Шаг 3	Диагноз	0	375	120	75,8
		1	69	190	73,4
	Общая процентная доля				74,9
Шаг 4	Диагноз	0	384	111	77,6
		1	64	195	75,3
	Общая процентная доля				76,8
Шаг 5	Диагноз	0	388	107	78,4
		1	59	200	77,2
	Общая процентная доля				78,0

Повторяя процедуру логистической регрессии, построим модель с новой точкой отсечения 0,332. Получится новая модель, которая обладает следующей предикторной способностью: чувствительность – 77,2%, специфичность – 78,4%, общее качество модели – 78%. Таким образом модель улучшилась на 0,4%.

Каждый раз, повторяя процедуру логистической регрессии, новый порог отсечения менял принадлежность людей к двух классам, а именно: с наличием заболевания и отсутствием заболевания.

Чтобы выбрать оптимальный порог классификации для данной задачи нужно сравнить площадь под ROC-кривыми для каждой точки отсечения:



Кроме графика ROC-кривой, для оценки качества регрессионной модели используется такая характеристика, как площадь под кривой (AUC). Можно считать, что чем больше показатель AUC, тем лучшей предсказательной способностью обладает модель.

Переменные результата проверки	Область
PGR_1	0,725
PGR_2	0,684
PGR_3	0,772
PGR_4	0,778

Из данной таблицы мы видим, что площадь под кривой с точкой отсечения 0,138 (0,684) говорит нам о среднем качестве модели; площади под кривыми с точками отсечения: 0,5 (0,725), 0,328 (0,772), 0,332 (0,778) – о хорошем качестве модели.

Сравнивая общее качество модели логистической регрессии и площади под кривыми, можно сделать вывод, что оптимальным порогом классификации является точка 0,332, которая обеспечивает максимум чувствительности и специфичности теста.

Заключение

В теоретической части было рассказано о таких темах, как логистическая регрессия; ROC-анализ.

В практической части было изучено построение модели логистической регрессии и применение метода ROC-анализа для улучшения и оценки качества полученной модели в задаче диагностики сахарного диабета. Для этого был использован прямой шаговый метод отношения правдоподобия для создания модели логистической регрессии. Исследование дополнено ROC-анализом и построением ROC-кривой.

Предикторная способность модели была улучшена с помощью выбора оптимальной точки отсечения. Также были определены: чувствительность, специфичность, общее качество модели и площадь под ROC-кривой для выбора наилучшей модели логистической регрессии.

Список литературы

1. Григорьев, С.Г. Роль и место логистической регрессии и ROC-анализа в решении медицинских диагностических задач / С.Г. Григорьев, Ю.В. Лобзин, Н.В. Скрипченко // Журнал инфектологии. – 2016. – Т. 8, № 4. – С. 36 – 45.
2. Паклин, Н.Б. Бизнес-аналитика: от данных к знаниям (+CD): Учебное пособие / Н.Б. Паклин, В.И. Орешков. – 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.
3. Наследов, А. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных / А. Наследов. – СПб.: Питер, 2013. – 416 с.
4. Филиппова, И.А. Прогностическое значение ранней диагностики сахарного диабета / И.А. Филиппова // Ремедиум. – 2014. – № 3. – С. 40 – 41.
5. https://ru.wikipedia.org/wiki/Кровяное_давление
6. https://ru.wikipedia.org/wiki/Индекс_массы_тела
7. <https://basegroup.ru/community/articles/logis-medic-scoring>