Семинар 6-7: Удержание клиентов, метрики классификации

Задача 1

Бандерлог оценил три модели: нейросеть, случайный лес и KNN. Он построил на тестовой выборке прогнозы и получил три матрицы ошибок:

	$y = 1$	$y = 0$
$\hat{y} = 1$	80	20
$\hat{y} = 0$	20	80

	$y = 1$	$y = 0$
$\hat{y} = 1$	48	2
$\hat{y} = 0$	52	98

	$y = 1$	$y = 0$
$\hat{y} = 1$	10	20
$\hat{y} = 0$	90	10000

а) Найдите для всех трёх моделей долю правильных ответов. Чем плоха эта метрика?

б) Найдите для всех трёх моделей точность (precision) и полноту (recall)

в) Предположим, что целевая переменная y принимает значение 1, если заемщик вернул кредит и 0, если не вернул. Вы хотите научиться прогнозировать платежеспособность клиента. Какую из первых двух моделей вы бы выбрали в таком случае?

г) Предположим, что целевая переменная y принимает значение 1, если человек болен больной болезнью с болью и 0, если он здоров. Вы хотите спрогнозировать нужно ли человеку обследование. Какую из первых двух моделей вы б выбрали в этом случае?

Задача 2

Бандерлог из Лога[1] ведёт блог, любит считать логарифмы и оценивать логистические регрессии. С помощью нового алгоритма Бандерлог решил задачу классификации по трём наблюдениям и получил $b_i = \hat{P}(y_i = 1|x_i)$.

y_i	b_i
1	0.7
0	0.2
0	0.3

1. Постройте ROC-кривую.
2. Найдите площадь под ROC-кривой.
3. Постройте PR-кривую (кривая точность-полнота).
4. Найдите площадь под PR-кривой.
5. Как по-английски будет «бревно»?

Ещё задачи!

[1] деревня в Кадуйском районе Вологодской области

Задача 3

Бандерлог начинает все определения со слов «это доля правильных ответов»:

а) accuracy — это доля правильных ответов...

б) точность (precision) — это доля правильных ответов...

в) полнота (recall) — это доля правильных ответов...

г) TPR — это доля правильных ответов...

Закончите определения Бандерлога так, чтобы они были, хм, правильными.

Задача 4

Бандерлог обучил логистическую регрессию и получил вектор предсказанных вероятностей принадлежностей к классу 1.

y_i	b_i
1	0.9
0	0.1
0	0.75
1	0.56
1	0.2
0	0.37
0	0.25

а) Бинаризуйте ответ по порогу t и посчитайте точность и полноту для $t = 0.3$ и для $t = 0.8$.

б) Какой порог бы вы выбрали?

в) Постройте ROC-кривую и найдите площадь под ней.