Машинное обучение в экономике

Семинар 2. Метод ближайших соседей

Задание №1

У вас имеются данные о характеристиках клиентов нейтрального к риску банка.

Дефолт (Д)	1	0	1	0	0	1	0
Возраст (В)	40	10	50	30	20	25	40
Зарплата (З)	50	20	10	40	30	30	75
Выборка	Обучающая					Тестовая	

Возраст и заработная плата клиентов измеряны в годах и тысячах рублей соответственно. Для прогнозирования дефолта вы используете метод ближайших соседей с 3 соседями и расстоянием Манхэттен.

Если клиент успешно возвращает кредит, то банк зарабатывает 10 тысяч рублей. В противном случае банк теряет 15 тысяч рублей. Также, банк может застраховать клиента, что гарантирует отсутствие дефолта. Однако, с застрахованных клиентов банк зарабывает лишь 5 тысяч рублей.

1. Придумайте единицы измерения возраста, при которых исключение дохода из числа признаков не изменит прогнозов рассматриваемого метода.

2. Покажите в общем случае, что, при использовании расстояния Минковского в методе ближайших соседей, увеличение всех признаков на одно и то же число или в одинаковое положительное количество раз не приводит к изменению ближайших соседей.

3. Сделайте вывод о целесообразности осуществления стандартизации признаков в данном случае.

4. Исходя из содержательной формулировки задачи запишите цены различных типов прогноза.

5. Найдите оптимальный порог прогнозирования в случае, когда банку известна истинная условная вероятность дефолта.

6. Посчитайте прибыль прогнозов на обучающей и тестовой выборках.

7. Рассчитайте F1-метрику на обучающей выборке.

8. Оцените условную вероятность дефолта для последнего индивида из тестовой выборки, используя взвешенный метод ближайших соседей с теми же числом ближайших соседей и метрикой расстояния.

Задание №2

Нейтральная к риску фирма прогнозирует уход клиентов (1 - ушел, 0 - остался). С каждого оставшегося клиента фирма получает 5 тысяч рублей, а с ушедшего – ничего. Если фирма считает (на основании прогнозов модели), что клиент хочет уйти, то она дает ему 2 тысячи рублей и в таком случае клиент гарантированно остается.

Ваш кот запрыгнул на клавиатуру и случайно (а может и нет) безвозвратно удалил данные, включавшие 100 наблюдений. Однако, вы запомнили рассчитанные по этим данным метрики качества прогнозов модели:

$$\text{precision} = 0.25 \qquad \text{recall} = 0.2 \qquad \text{ACC} = 0.3 \qquad \text{FPR} = 0.6$$

Также, по удаленным данным у вас была составлена таблица с расчетами выигрыша (gain) и подъема (lift). При этом вы разбивали выборку не на 10 (децили), а по аналогии на 5 равных частей. Однако, помахав лапами, кот удалил часть значений в этой таблице.

Часть	Выигрыш	Кумулятивный выигрыш	Подъем	Кумулятивный подъем
1				1.5
2		0.56		
3	0.16			
4			0.8	
5				

Наконец, урчание счастливого кота заставило вас позабыть, чему равнялся AUC. Тем не менее вы помните, что $\text{pAUC}(0.5, 1) = 2\text{pAUC}(0, 0.5) = 0.5$. Также, у вас была построена еще одна модель, по поводу которой вы помните, что ее $\text{pAUC}(0.3, 0.7)$ был больше (чем у исходной модели), но AUC – меньше.

1. Рассчитайте прибыль от прогнозов модели на удаленных данных.
2. Восстановите значения, пропущенные в таблице.
3. Найдите значение AUC и нарисуйте любые ROC-кривые, удовлетворяющие вашим воспоминаниям о построенных моделях.