

Кейс. Задача кредитного скоринга

Контекст

Машинное обучение является одним из основных инструментов для прогнозирования вероятности дефолта по кредитам. Прогноз вероятности дефолта необходим для принятия практически всех решений, связанных с выдачей кредитных карт, установлением лимитов, а также для управления рисками. Также подобный анализ необходим для решений, связанных с выходом на новые рынки (и привлечением клиентов с уже существующих рынков) и с разработкой и предложением новых продуктов.

Данные

Ваш банк собрал данные по 24 000 клиентов. Из них 1 000 случайно выбранных клиентов банк «отложил» для пилотного проекта, описанного ниже. Данные по оставшимся 23 000 клиентов в файле «train_credit.xls». По каждому клиенту представлена демографическая информация (пол, образование, семейный статус и возраст), лимит кредитной линии, история по кредитной карте (статус платежа, сумма счёта, и сумма платежа) за последние шесть месяцев (с Апреля по Сентябрь), и наличие или отсутствие дефолта в Октябре. В таблице представлены данные по первым 20-ти клиентам:

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	default
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	689	0	0	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0	
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	48291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0	
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36581	10000	9000	689	679	0	
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0	
7	5.00E+05	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0	
8	1.00E+05	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0	
9	140000	2	3	1	28	0	0	0	2	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0	
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0	
11	2.00E+05	2	3	2	34	0	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0
12	260000	2	1	2	51	-1	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	2870	0	1	0
14	70000	1	2	2	30	1	2	2	0	0	0	65802	67369	65701	66782	36137	36884	3200	0	3000	3000	1500	0	1	0
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0	0
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0	0
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3300	0	1500	0	1650	0	1	0
18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0	0
19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Описание («словарь») данных

ID: Порядковый номер клиента

LIMIT_BAL: Размер кредитной линии (включая индивидуальные и семейные кредиты), в тысячах долларов.

SEX: Пол (1=мужской, 2=женский)

EDUCATION: Образование (1=мастер, 2=бакалавр, 3=школа, 4=другое, 5=неизвестно, 6=неизвестно) Порядковый признак (особая предобработка) !!!!!

MARRIAGE: Семейное положение (1=женат/замужем, 2=холост, 3=другое)

AGE: Возраст, в годах (Попробуйте категории возрастов)

PAY_1: Статус платежа в предыдущем месяце – в сентябре: (-2=не надо платить – нулевой баланс, -1=полностью оплачено, 0=использование возобновляемого/револьверного кредита, 1=задержка платежа на один месяц, 2=задержка платежа на два месяца, ... 8=задержка платежа на 8 месяцев, 9=задержка платежа на 9 месяцев и больше)

PAY_2: Статус платежа два месяца назад – в августе (шкала как описано выше для PAY_1)

PAY_3: Статус платежа три месяца назад – в июле (шкала как описано выше для PAY_1)

PAY_4: Статус платежа четыре месяца назад – в июне (шкала как описано выше для PAY_1)

PAY_5: Статус платежа пять месяцев назад – в мае (шкала как описано выше для PAY_1)

PAY_6: Статус платежа шесть месяцев назад – в апреле (шкала как описано выше для PAY_1)

BILL_AMT1: Баланс по карте в сентябре (в тыс. у.е.)

BILL_AMT2: Баланс по карте в августе

BILL_AMT3: Баланс по карте в июле

BILL_AMT4: Баланс по карте в июне

BILL_AMT5: Баланс по карте в мае

BILL_AMT6: Баланс по карте в апреле

PAY_AMT1: Размер платежа в сентябре (в тыс. у.е.)

PAY_AMT2: Размер платежа в августе

PAY_AMT3: Размер платежа в июле

PAY_AMT4: Размер платежа в июне

PAY_AMT5: Размер платежа в мае

PAY_AMT6: Размер платежа в апреле

Default_0: Наличие дефолта в Октябре (1=да, 0=нет)

Задача:

Ваш отдел планирует новую акцию по предоставлению краткосрочных кредитов без залога на сумму 25,000, и для цели этого задания допустим что всего на 1 месяц под 2% в месяц. Более того, каждый клиент, получивший и выплативший кредит, будет с большей вероятностью обращаться в Ваш банк за последующими кредитами, что, по Вашим прогнозам, имеет дополнительную пожизненную ценность клиента (customer lifetime value, CLV) в 1,000.

Данные по 1 000 клиентов, которые случайно выбраны для пилотного проекта новой акции, представлены в файле «test_credit.xlsx». В таблице представлены данные по первым 5-ти клиентам:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
1	ID	LIMIT	BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
2	n1000-1	500	1	1	2	29	0	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770
3	n1000-2	210	1	1	2	29	-2	-2	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0
4	n1000-3	150	1	1	2	27	0	0	0	0	0	0	0	86009	86108	89006	89775	87725	40788	4031	10006	3266	4040	1698	800
5	n1000-4	20	1	2	1	38	0	0	0	0	0	0	-1	17973	19367	19559	18240	17928	150	1699	1460	626	1750	150	0
6	n1000-5	400	1	2	1	34	-1	-1	-1	-1	-1	-1	-1	19660	9666	11867	7839	14837	7959	9677	11867	7839	14837	7959	5712

Кому из этих клиентов Вашему отделу следует выдать кредиты?

Прим. Для простоты задания сделайте следующие допущения:

- (i) дефолт по уже выданным кредитам - не ваша проблема.
- (ii) в случае дефолта через процедуру взыскания долгов можно вернуть 20,000 из 25,000 выданного кредита
- (iii) все клиенты, которым будет предложен кредит, возьмут его и на полную сумму в 25,000
- (iv) стоимость капитала на 1 месяц = 0.

Иными словами, для каждого клиента: если кредит не выдан, то прибыль=0. Если выдан и возвращен, то прибыль = $2\% \times 25,000 + 1000 = 1,500$. Если выдан и не возвращен, то убыток в 5,000.

Как будет проверяться контрольная работа?

Перейдите по данной ссылке:

https://docs.google.com/forms/d/108qNFGtG9ZMeWV4nUGHTR2JXKDcJXlZFeGIQighcDMk/viewform?edit_requested=true

1) заполните ФИО [0 баллов]

2) скопируйте и вставьте свое решение (1 выдать, 0 не выдать)

Посчитается прибыль, которую Банк на самом деле получит, следуя вашей рекомендации [Есть данные, кто из этих 1000 клиентов на самом деле вернет кредит а кто нет] и будут выставлены баллы по следующей шкале:

- a. если Ваша прибыль ≤ 0 , 0 баллов
- b. если Ваша прибыль > 0 но $\leq 100,000$, 3 балла
- c. если Ваша прибыль $> 100,000$ но $\leq 300,000$, 5 баллов
- d. если Ваша прибыль $> 300,000$ но $\leq 500,000$, 10 баллов
- e. если Ваша прибыль $> 500,000$, 15 баллов

3) по каким трём клиентам Вы наиболее уверены в том, что они вернут кредит?

4) по каким трём клиентам Вы наименее уверены в том, что они вернут кредит?

Многие модели машинного обучения «соглашаются» по «лучшим» и «худшим» клиентам. Так дело обстоит и в этом случае. Вы получите по 3 балла за вопросы №3, №4 если 2 или более выбранных вами клиента указаны правильно (т.е. на самом деле вернут или не вернут кредит, в соответствии с условием задания)

Максимальный балл = 21.