Модель оценки новых (первичных) клиентов для региона “Вьетнам”

21.01.2020

Процесс создания модели

- Модель основана на данных 2-х организаций – Kalapa (кредитный сервис) и Trusting Social (кредитный сервис) и клиентской информации, указываемой при заполнении заявки клиентом по умолчанию.

- Имелась выборка размером 1420 наблюдений (все новые клиенты, получившие заем) за период от 01.01.2019 до 20.10.2019.

- В качестве таргета использовались клиенты, которые за 60 дней с момента выдачи займа не отдали нисколько денег (т.к. при исключении такого рода заявок мы бы получали наибольшую доходность по сегменту новых клиентов). Получаем разметку выборки (таргет): 1 – если клиент не отдал нисколько денег за 60 дней со дня выдачи, в противном случае – 0. Соотношение: 1 – 47%, 0 – 53%.

- Выборка была разделена случайным образом на 2 части - 1065 и 355 строк, но с сохранением соотношения (47% на 53%) значений таргета (1 и 0) в обоих частях выборки. 1-я часть выборки (1065 строк) была использована для обучения алгоритма МО (т.е. для извлечения паттернов из выборки, соотнося значения признаков с разметкой таргета), а 2-я часть выборки использовалась для независимого тестирования и проверки качества работы модели обученной на 1-й части.

- Для обучения использовался алгоритм градиентного бустинга случайного леса (XGBoost), т.к. алгоритм показал наилучшие результаты (максимальное значение метрики ROC-AUC на 2-й части выборки) в сравнении с другими популярными алгоритмами МО (линейная регрессия, логистическая регрессия, случайный лес).

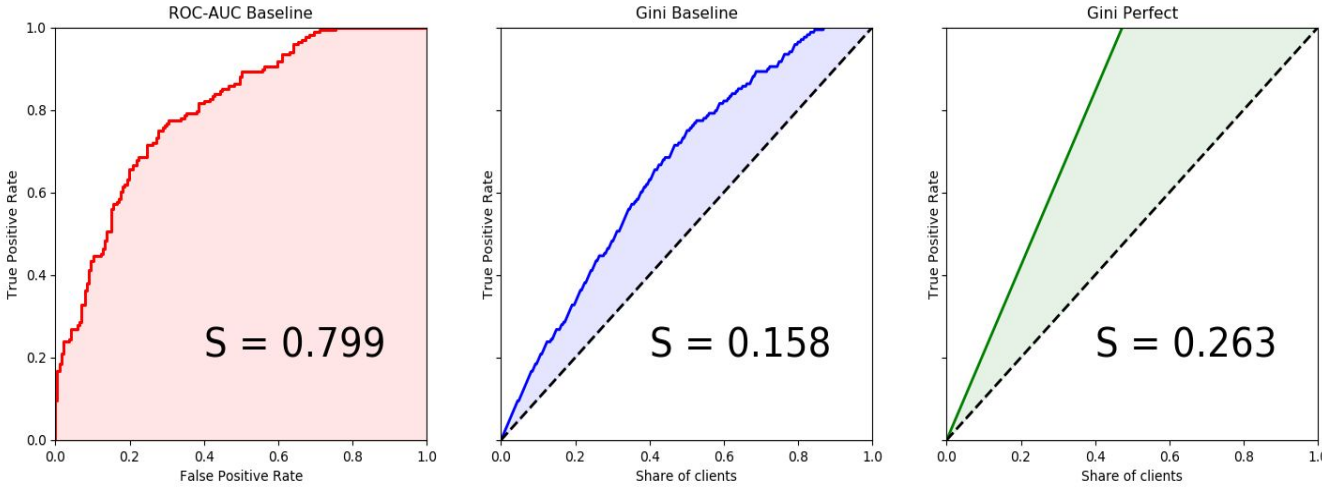
- На выходе мы получили математическую модель, которая способна принимать максимально верное решение по клиенту, исходя из тех данных (исходная выборка), которые мы имели, при этом обрабатывая пропуски в присылаемых данных, если таковые имеются.

Используемый вектор признаков в модели, и их приоритет при обучении (информативность признаков)

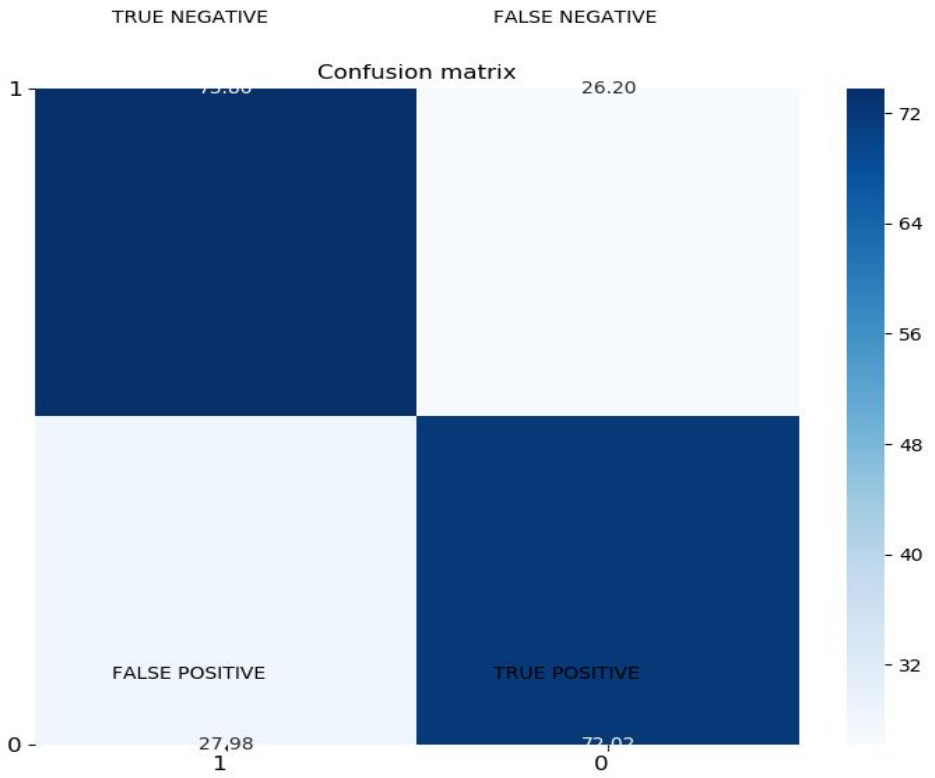
KC_ - признаки кредитного сервиса Калапа (2шт). Топ 2 самых “важных”
TS_ - признак кредитного сервиса Трастинг Сошиал (1шт). 3 место по “важности”
CU_ - признаки указываемые самим клиентом по умолчанию (9шт). Признаки оказывающие остаточное влияние на решение.

Значения некоторых основных метрик оценки качества модели на тестовой выборке (2-я часть выборки)

Gini = 2*AUCROC-1 = 0.5980392156862742



Площадь под ROC-кривой, Gini



Матрица ошибок

Результаты работы модели на тестовой выборке

Ранжирование балла	Количество клиентов	% клиентов	Доходность	Средний доход	Средняя просрочка	
0,05-0,1	9	2,5%	138,7%	524 390	6	47,3% заявок (одобрение)
0,1-0,15	46	13,0%	136,6%	456 517	12	
0,15-0,2	18	5,1%	95,0%	4 770	77	
0,2-0,25	12	3,4%	113,0%	46 083	77	
0,25-0,3	15	4,2%	120,1%	204 550	62	
0,3-0,35	26	7,3%	96,1%	-85 212	78	
0,35-0,4	19	5,4%	99,3%	-102 526	79	
0,4-0,45	23	6,5%	101,1%	37 163	67	
0,45-0,5	17	4,8%	79,3%	-269 279	118	52,7% заявок (отказ)
0,5-0,55	22	6,2%	70,5%	-269 072	74	
0,55-0,6	22	6,2%	60,2%	-436 250	111	
0,6-0,65	31	8,7%	44,8%	-829 669	159	
0,65-0,7	26	7,3%	33,8%	-793 327	156	
0,7-0,75	14	3,9%	48,6%	-656 339	165	
0,75-0,8	13	3,7%	67,7%	-271 577	110	
0,8-0,85	15	4,2%	28,0%	-893 333	158	
0,85-0,9	26	7,3%	5,4%	-1 196 154	193	
0,9-0,95	1	0,3%	0,0%	-1 000 000	114	
Общий итог	**355**	**100,0%**	**78,1%**	**-277 585**	**99**	

Модель проставляет значение (“балл”) от 0 до 1, которое обозначает вероятность принадлежности к классу таргета равному 1 (клиент, размеченный как 1 – клиент, который за 60 дней с момента выдачи займа не платил денег), т.е. чем меньше “балл”, тем меньше вероятность того, что клиент будет принадлежать к классу 1. При делении клиентов на группы по “баллу” мы наблюдаем практически линейный рост доходности от групп с большим “баллом” к меньшим.