












«Прогнозирование вероятности невозврата кредита»

сроки: 01.08 – 01.12 / 2017

Организаторы: САС Институт
Хоум Кредит банк

sascompetitions.ru

Топ 5 из ~130

1	 Anzor Berezgov
	 s20_940-2.csv 0.717437
2	 Ivan Timoshilov
	 0.712104
	HC_11_30_3.csv
3	 alexanderdyakonov
	 ridge_C04.csv 0.71172
4	 IzmaylovKK 0.711398
	 uio-1.csv
5	 Паша Коваленко
	(MMP, MSU, Russia) 0.710919
	 lstm_v10_augm_final.csv

Данные

Информация из 4х бюро кредитных историй:

TRAIN:

1 787 571 записей

135 155 клиентов

DEF.mean = 3.35%

TEST:

1 665 298 записей

120 567 клиентов



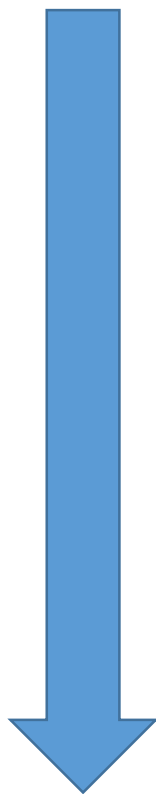
Задача: оценить DEF для TEST

Метрика: ROC AUC

Public: «часть тестовой выборки», Private: «вся тестовая выборка»

Данные

В
р
е
м
я



Кредит 1	1
Кредит 2	1
Кредит 2	2
Кредит 3	3
Кредит 3	3
Кредит 3	4



Клиент

Данные

ID – идентификатор заявки

SK_DATE_DECISION - дата рассмотрения заявки

NUM_SOURCE - номер источника данных

CREDIT_ACTIVE - статус кредитного договора (0 - договор закрыт, 1 - договор активен, 2 - договор продан, 3 - безнадежный долг)

CREDIT_CURRENCY валюта кредита

DTIME_CREDIT - дата выдачи кредита

CREDIT_DAY_OVERDUE - текущая просрочка, дни

DTIME_CREDIT_ENDDATE -планируемая дата окончания кредита

DTIME_CREDIT_ENDDATE_FACT - фактическая дата окончания кредита

AMT_CREDIT_MAX_OVERDUE - максимальная просроченная задолженность за все время жизни договора, сумма

CNT_CREDIT_PROLONG - число пролонгаций кредита

AMT_CREDIT_SUM - сумма кредита

AMT_CREDIT_SUM_DEBT - сумма оставшегося долга

AMT_CREDIT_SUM_LIMIT – лимит (для карт)

AMT_CREDIT_SUM_OVERDUE - текущая просроченная задолженность, сумма

CREDIT_TYPE - тип договора(0 - неизвестный тип кредита, 1 - кредит на автомобиль, 2 – лизинг,

3 – ипотека, 4 - кредитная карта, 5 - потребительский кредит ...)

DTIME_CREDIT_UPDATE - дата последнего обновления информации в источнике

CREDIT_DELAY5 , 30, 60, 90, MORE - число просроченных на N дней платежей

AMT_REQ_SOURCE_HOUR, WEEK, MON, QRT, YEAR - число запросов к источнику за последний час

AMT_ANNUITY - сумма ежемесячного платежа

TEXT_PAYMENT_DISCIPLINE - платежная строка, вектор статусов платежей по кредиту

33 поля + DEF

Данные

TEXT_PAYMENT_DISCIPLINE платежная строка, вектор статусов платежей по кредиту, возможные статусы:

0 – своевременный платеж,

1,2,3,4 – просрочка 1..30, 31..60, 61..90, 91..120,

5 - просрочка 121+ дней, передан коллекторам, продан, списан

X – статус неизвестен

С – договор закрыт

Пример хорошего заемщика: CCCCCCCCCCCCCC00000000000000000000000000000000

Процесс

Delphi:

Предобработка и генерация признаков, постобработка

Python:

обучение модели, предсказание

Обработка данных

Возможные варианты:

- 1) Аггрегировать несколько записей по одному кредиту от разных БКИ в одну
- 2) Попытаться заполнить недостающие данные
- 3) Разрешить противоречия между записями

Решение: ничего не делать

Подход №1

Каждая запись по клиенту рассматривается как отдельная независимая сущность.

Для предсказания выбирается максимум или средний скор по всем записям клиента.

ROC AUC ~ 68-69

Подход 2

Соединение N последних записей клиента в один вектор

+ одна агрегация по всем кредитам

На усреднении подходов 1 и 2 : ROC AUC \sim 70

Подход 3

Агрегация кредитов клиента в различных срезах
+ комбинация с подходом 1 в соотношении 4:1

ROC AUC = 72.4
(финальная модель)

Признаки

Группировки кредитов по DTIME_CREDIT

12 групп:

0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 9 лет

Признаки

Каждая группа рассматривалась в разрезах (10 шт):

активность: Активные, неактивные,

тип кредита: Все типы,

1 - автокредит, 4 – кредитная карта,

5 - потребительский, 19 – микрозайм

Итого: 120 подгрупп + 1 подгруппа без разрезов

Признаки

Каждая из 121 групп содержит

Базовый блок из 147 признаков

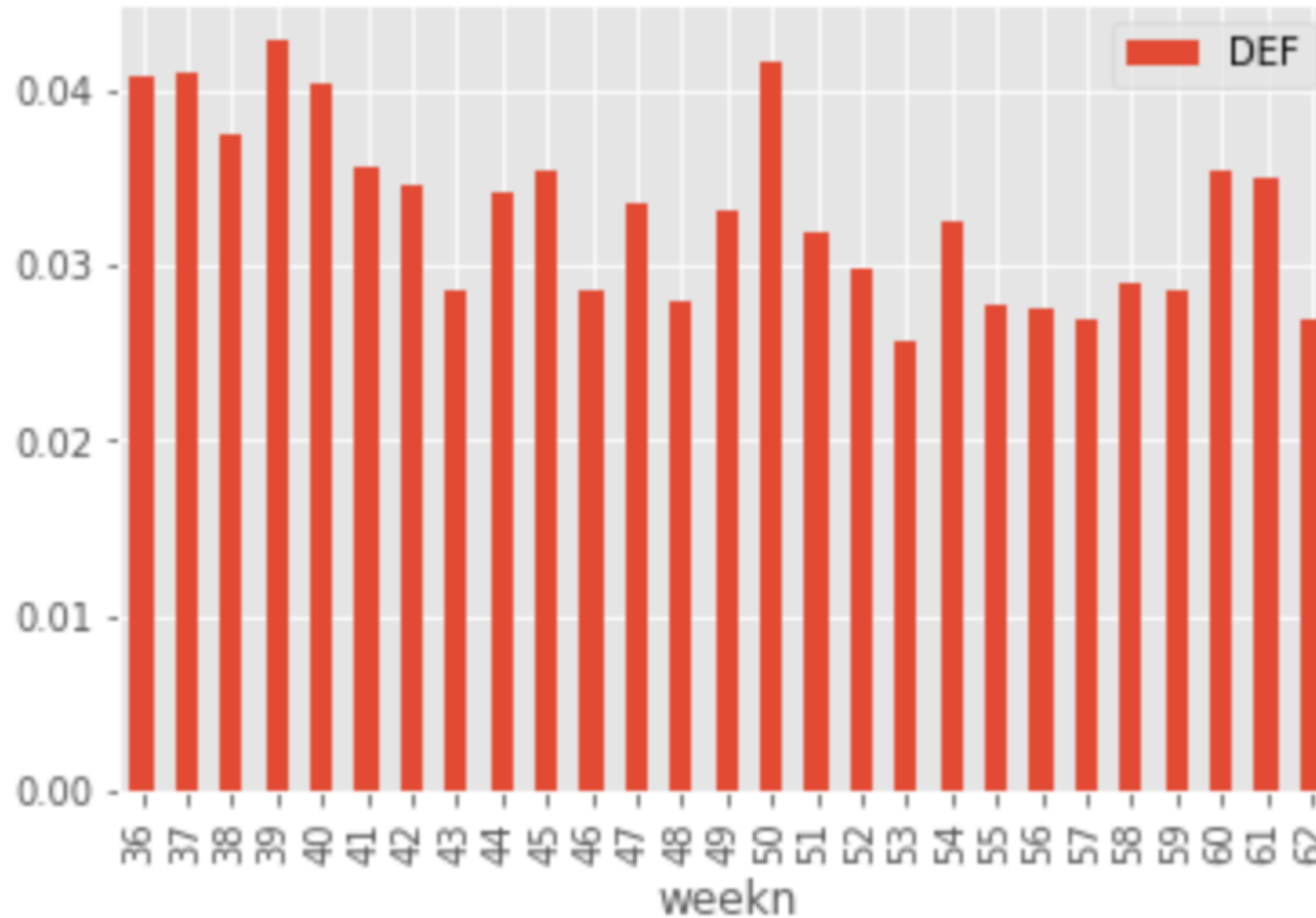
Итого на группах создано: 17 787 признаков

+ использованы 9 предыдущих кредитов

(по 42 признака на кредит) = 378 признаков

+ ID

Зависимость DEF от weekN



Признаки

Всего признаков: 18166

Сразу перед обучением модели
выброшено 4403 признака, по количеству
уникальных значений ≤ 1

Для обучения модели: 13 764 признака.

Признаки (147 шт)

COUNT (категориальных признаков)

- по типам кредита
- по источникам
- по числу обращений в БКИ
- по числу просрочек

STD, AVG, MIN, MAX, SUM (числовых признаков)

числовых переменных

Признаки

**STD, AVG, MIN, MAX, SUM – всех
ВОЗМОЖНЫХ ЧИСЛОВЫХ переменных**

TEXT_PAYMENT_DISCIPLINE - 80 признаков

Количество символов каждого вида

+ срезы строки за 1, 2, ..., 9 мес по отдельности

+ срез за последний год

Использованные алгоритмы

- 1) CatBoost - первоначально
- 2) Lightgbm – финальная модель

Подбор гиперпараметров вручную

Параметры Lightgbm

Основные параметры:

num_leaves: 68

feature_fraction: 0.7

bagging_fraction: 1

max_depth:-1

Параметры Lightgbm

Борьба с оверфитом:

reg_alpha: 5

reg_lambda: 11

min_split_gain: 0.5

min_data_in_leaf: 15

min_sum_hessian_in_leaf: 1

Параметры Lightgbm

Примерное число деревьев рассчитано на
кросс-валидации + 90 деревьев
(подобрано по лучшему скору на лидерборде)
Итого деревьев: **940**

Время обучения ~4 ч.

Железо

Ноутбук:

Intel Core i7-4500U + 8Gb

Стационарный комп:

Intel Core i9-7980XE + 128Gb

Немного статистики

87 различных моделей

180 сабмитов

427 строк на Python

1969 строк на Delphi

Спасибо за внимание!