ML Boot Camp V: Предсказание сердечно-сосудистых заболеваний

ШАЯХМЕТОВ РИМ

Постановка задачи

- 100 000 пациентов. Train 70%, Public 10%, Private 20%.
- Предсказание вероятности CC3 по результатам врачебного осмотра.
  - Возраст, рост, вес, пол
  - Верхнее и нижнее давление. Уровень холестерина и глюкозы (3 категории)
  - Курение, алкоголь, физическая активность (бинарные признаки)
    - 10% скрыто на тестовых данных
- Метрика - logloss

Данные

train

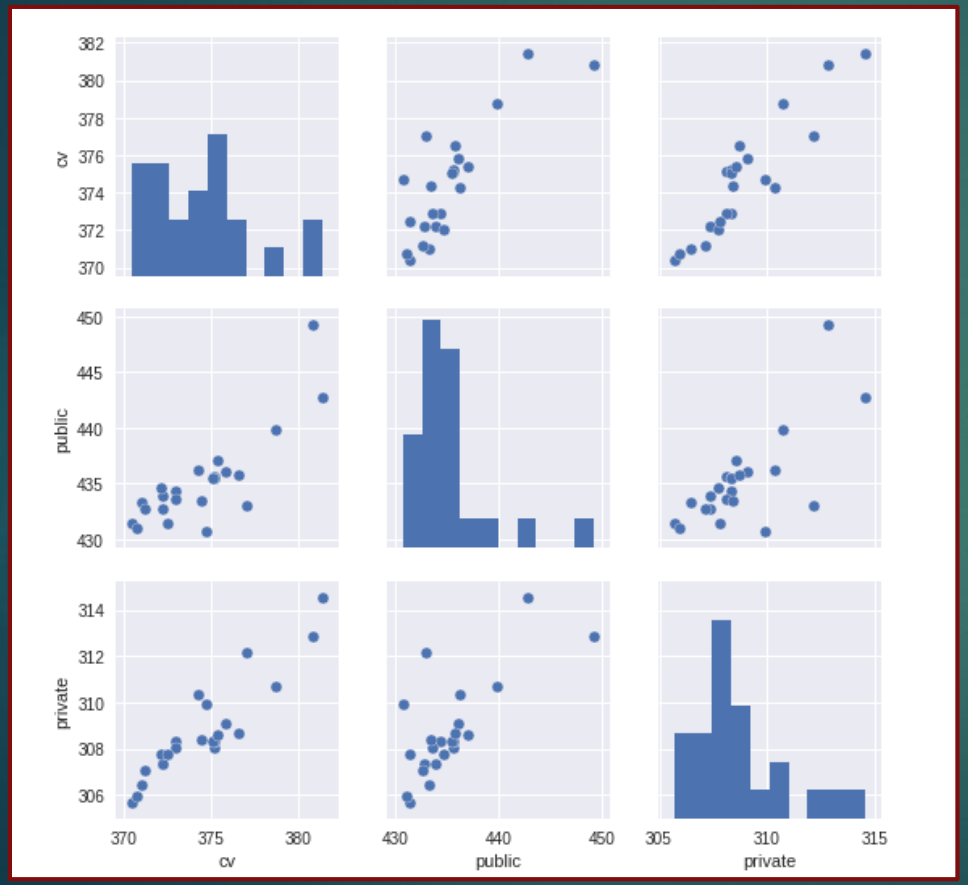
	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0
5	8	21914	1	151	67.0	120	80	2	2	0	0	0	0

- Возраст в днях
- Сбалансированная выборка
- Грязные поля – давление (ap_hi, ap_lo), рост, вес

Кросс-валидация

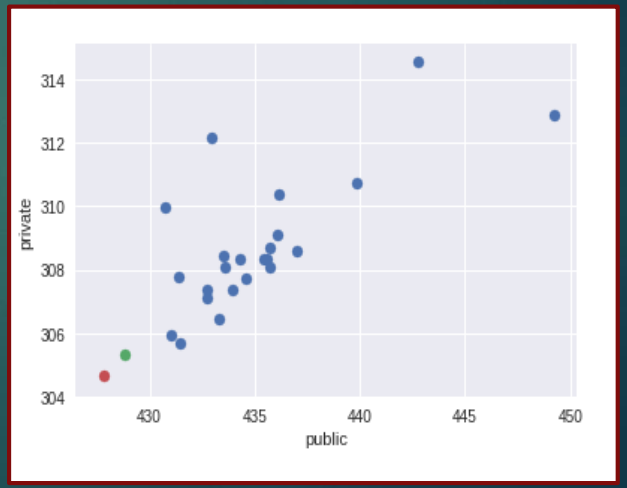
- 7 CV (10 000 в валидации).
- В валидации 10% alco, smoke, active скрыты
- Несколько стратегий по изменению alco, smoke, active
  - Оставить как есть (NaN только в валидации)
  - Скрыть 10% в обучении (алгоритм должен принимать NaN, пр. XGB)– почти всегда показывал лучший CV.
  - Предсказать в валидации – для алгоритмов, которые не умеют работать с NaN
  - Предсказать вероятности для всех значений и заменить на них – совсем не сработало
- Если использовать стандартную CV без NaN, то результаты были лучше, но, видимо, завышенными.

Корреляция с лидербордом

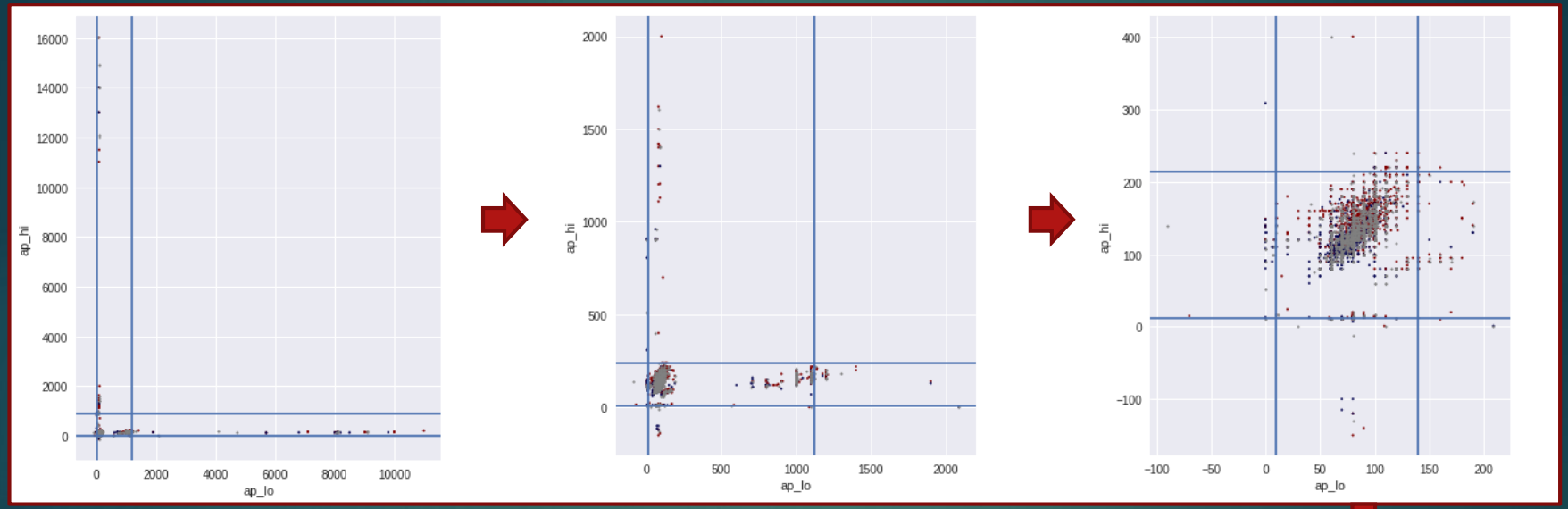


Spearman rho correlation

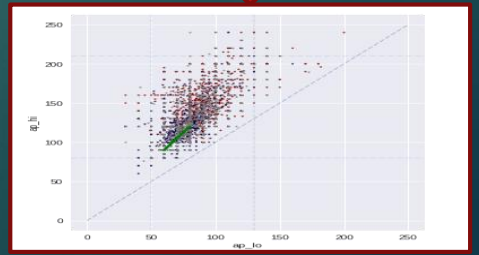
	CV	Public	Private
CV	1	0.723	0.915
Public		1	0.643
Private			1



Чистка данных I



- Улучшало результат на лучших моделях до CV ~0.5375, Public ~0.5435
- Аналогично с ростом-весом



Модели

- XGBoost - почти все наилучшие модели, включая усреднение различных xgb
- Neural Networks (Keras) - пробовал, но не смог в итоге улучшить результат xgb
- Усреднение xgb с другими моделями (RF, NN, ExtraTrees) не давало улучшения cv
- Стекинг (brew) тоже не улучшал результат усреднения различных xgb

➡ Сконцентрировался больше на очистке данных, преобразованиях признаков и поиске 1-3 оптимальных xgb

Гиперпараметры искал с помощью Байесовской оптимизации с большим количеством случайного поиска в качестве инициализации

Чистка данных II

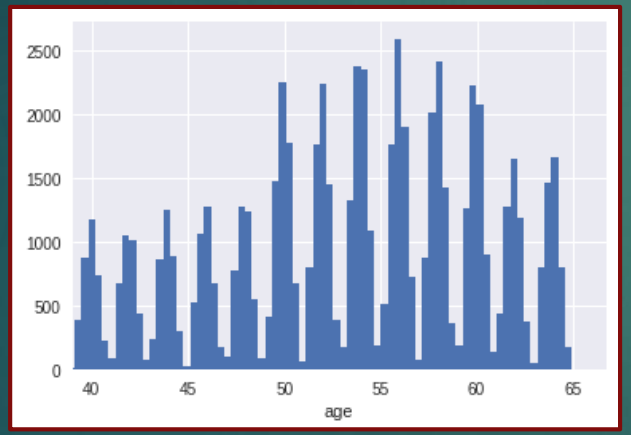
- Основная идея – к каждому правилу есть исключение
  - Делить на 10 у давлений между 1000-2000, но у давлений 1130, 1420, 1620 игнорировать последние две цифры (110, 140, 160)
- Смотреть на другие поля
  - Давление в 12000, 13000 делим на 100, но давление 14900/90 - это скорее всего 140/90
  - 1/1099, 1/2088 заменить на 110/90 и 120/80
  - Самые сложные случаи – замена 585 на 85, 701 на 170, 401 на 140
- Находить похожие в train в более неоднозначных ситуациях
  - 13/0 заменить на 130 на 80, так как более вероятно
- Помнить, что встречаются аномальные реальные значения (150/60 лучше оставить, тем более если это в train и CC3)
- Рост\вес очищался по аналогичным принципам, но большинство неоднозначностей так и остались неразрешимыми

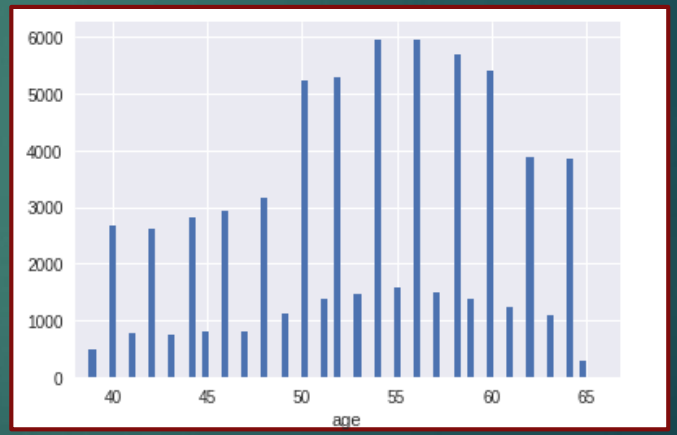
Чистка данных II

- По недавнему выложенному полному датасету чистка давления, роста и веса в итоге оказалась проделанной
  - Для 1379 объектов в train
  - Для 402 объектов в private
  - Для 194 объектов в public
- Все последующие модели с данной чисткой смогли улучшить CV до ~0.5370 (с ~0.5375), Public до ~0.5431 (с ~0.5435)
- Правильность чистки не 100%, но помогла значительно улучшить результат
- Желательно после каждой чистки, преобразований находить более оптимальные гипер-параметры xgb (помимо тривиального количества деревьев)

Преобразование признаков I

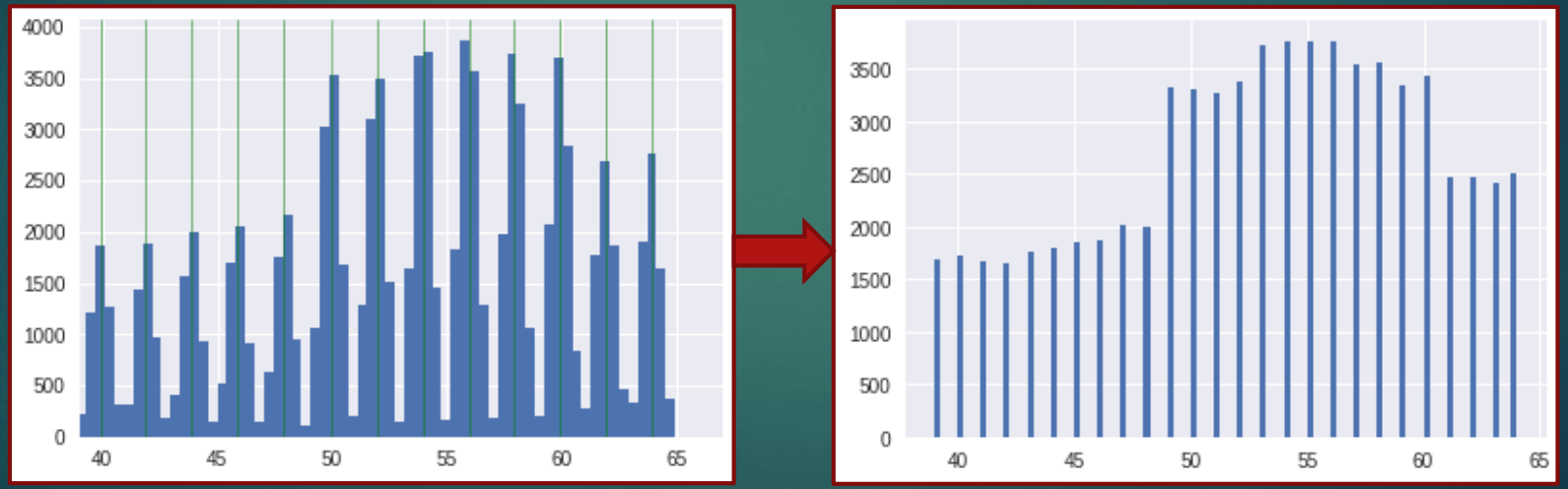
- Возраст
  - Если поделить количество дней на 365.25, то можно получить распределение по годам, но:





Преобразование признаков I

- Улучшая CV и гистограмму, разбиваем возраст по «годам», соответствующим половине распределений в гауссовской смеси



Новые признаки

- Поиск и отбор проводился вручную
- Комбинации более чем двух признаков всплывали в feature importance, но ухудшали CV
- BMI = weight/(height/100)**2 -> первый в feature importance
- Pulse pressure = ap_hi – ap_lo
- Нормальное ли давление

  (85 <= ap_hi <= 125 & 55 <= ap_lo <= 85)
- Последняя цифра в давлении (+перестановка)
- Аналог чётности года (age/365.25) = (age – (age/2).round()*2) > 0

Преобразование признаков II

- Дискретизация признаков (bmi, weight, height)
  - Почему? Возраст разделял лучше 0 от 1 (roc auc = 0.6358), чем BMI (roc auc = 0.6151), когда как в моих моделях BMI имел значительно высокий feature importance (может это и нормально)
  - Зачем? Улучшило CV моделей (1-3 xgb) + полезно для смешивания с моделями на исходных признаках
- Порог дискретизации на основании квартилей, количество на основании cv + хорошие графики + feature importance
- Округление давления, пульса

Последний час соревнования

- До конца соревнования чуть больше часа:
  - Все лучшие модели были простые (1-3 xgb, простое усреднение, либо с весами) на основе последней чистки данных, преобразований (включая дискретизацию).
  - Усреднение 2 xgb с применением вышеуказанных приёмов – github.com/shayakhmetov/mlbootcampV (CV 0.5370, Public 0.5431, Private 0.530569 = 2 место)

Последний час соревнования

- Усреднение 9 предсказаний с весом, обратно пропорциональной округлённой 4 цифре на Public: улучшение Public с 0.5431 до 0.54288
- Добавив ещё более различные предсказания (с ещё меньшими весами), взвешенное усреднение 17 предсказаний дало Public 0.54278
- Переобучение?
  - Больший вклад от моделей со стабильным CV около 0.5370-0.5371
  - Большинство моделей использовало последнюю чистку данных
  - Модели различались дискретизацией (отсутствием), преобразованием с возрастом, дополнительными признаками, разными seeds, стратегиях с NaN, т.д.
- Итог: 0.5304688 Private = 2 место

Выученные уроки

- Использование относительно простых моделей на разных признаках/предобработке может дать лучшие результаты, чем использование многих моделей на одних и тех же данных
- Необходимо хранить результаты кросс-валидации промежуточных моделей, чаще коммитить в git. Иначе, взвесив множество предсказаний (и не переобучившись), не совсем понятно, комбинация каких именно признаков\моделей дала прирост. (к любому правилу исключение – если только не осталось одного часа до конца соревнования)
- Нужно было продолжать эксперименты с добавлением нейронных сетей
- Стекинг?

Спасибо за внимание!