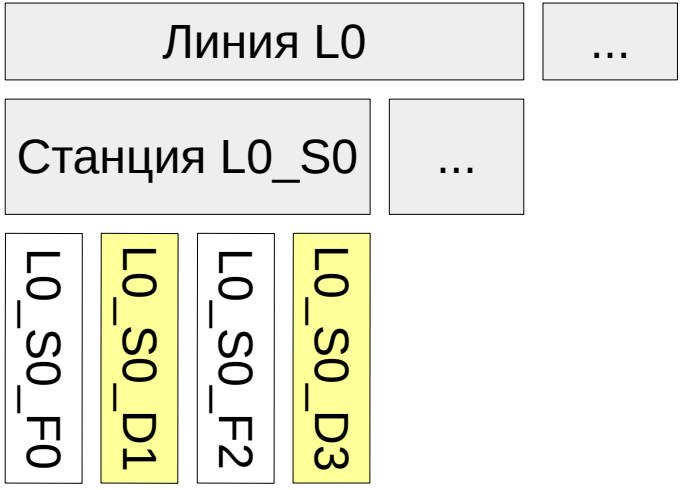
Bosch Production Line Performance

Алексей Носков

Задача

Прогнозирование внутренних сбоев в производстве

- Деталь движется по линиям
- Проходит через станции
- На станциях снимаются замеры



Задача

2.37M примеров:

- 50% - train
- 15% - public test
- 35% - private test

4268 признаков:

- 970 числовых
- 1157 временных (в неизвестных единицах)
- 2141 категориальных

Задача

Мало положительных примеров - 0.6%

Метрика: Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Команда

- Gilberto Titericz Junior
- Michael Jahrer
- DataGeek
- Alexey Noskov

Организация работы

- Координация в Slack
- Общий файл с CV-разбиением
- Файлы с прогнозами моделей
- Файлы с фичами

Подготовка данных

- Удаляем столбцы с единственным значением
- Удаляем столбцы-дубликаты

После удалений:

- 927 числовых
- 161 временная
- 227 категориальная

Базовые признаки

- Среднее, минимум, максимум по станциям
- Кол-во станций
- Начальная/конечная станция
- Кол-во нулей в числовых столбцах
- Кол-во непропущенных значений

Даты

Предположительно: 0.01 = 6 минут (beluga)

Есть суточные и недельные циклы

Признаки:

- Время прохождения линии
- Количество выходных во время прохождения линии
- Среднее время для станций линии

Даты

Большая нагрузка приводит к сбоям?

- Кодируем даты количеством примеров

Сбой на станции затрагивает несколько примеров?

- Кодируем даты количеством сбоев
- Суммируем по линиям

Порядок и дубликаты

Некоторые соседние строки практически идентичны

5537	-0.062	-0.064	-0.161	-0.179	-0.056	0.07	0
5541	0.101	0.063	-0.015	0.003	-0.013	0.116	-0.015
5542	0.101	0.063	-0.015	0.003	-0.013	0.116	-0.015
5544	0.101	0.063	-0.015	0.003	-0.013	0.116	-0.015
5545	0.043	0.123	0.003	-0.016	0.031	0.07	-0.007

В дублирующихся строках сбои бывают чаще!

Порядок и дубликаты

Природа пока неизвестна:

- Утечка при подготовке данных?
- Реальная особенность?

Vicens Gaitan: *После сбоя в обмене данными SCADA повторяет последние известные значения*

Порядок и дубликаты

Бывают более сложные структуры:

6310	-0.014	…	-0.027	0.002				0	0	0	0	-0.195	-0.319
6311	-0.014	…	-0.027	0.002	-0.006	0.037	0.098						
6312	-0.014	…	-0.027	0.002	-0.006	0.037	0.098						
6313	-0.014	…	-0.027	0.002	-0.002	-0.113	-0.003						
6314	-0.014	…	-0.027	0.002	-0.002	-0.113	-0.003						
6315	-0.014	…	-0.027	0.002	-0.002	-0.15	-0.015						
6316	-0.014	…	-0.027	0.002	-0.002	-0.163	0.033						

Порядок и дубликаты: признаки

Скрипт Faron:

- Сортируем по минимальной дате
- Разница id со следующей/предыдущей строками

Дополнения:

- Факт сбоя в соседних примерах
- Факт отличия максимальной даты в соседних строках

Порядок и дубликаты: признаки

Оценка похожести соседей:

- Кол-во столбцов, отличающихся в соседних строках

Информация о группе дубликатов:

- Размер группы
- Позиция строки в группе

Алгоритмы

- Xgboost
  - 2k итераций, глубина 9
- Нейронные сети (mxnet, keras)
  - 2-3 скрытых слоя
- Random Forest, Extremely Randomized Trees

Xgboost

Отбор признаков:

- Загружаем небольшое подмножество данных
- Проводим 10-100 итераций бустинга
- Выбираем наиболее значимые признаки
- Загружаем полный набор данных только с этими признаками

Кросс-валидация

Stratified K-Fold

- 5 folds

Выбор порога для MCC на out-of-fold прогнозах

Начиная с 0.49х работала плохо

Стекинг

На втором уровне

- Xgboost
- Нейронные сети

Обучение на out-of-fold прогнозах первого уровня

Кросс-валидация на том же разбиении

Усреднение нескольких моделей второго уровня

Результаты

Private LB: 6 место

Public LB: 3 место

Мы заоверфитили Public LB, большинство изменений в последние недели не несло улучшений

Вычислительные ресурсы

Ноутбук 4 потока, 8GB свободной памяти

AWS EC2 Spot instances

- 8-16 ядер в зависимости от стабильности цен, 0.1-0.3 $/час
- Цены очень нестабильны

Альтернатива: Google Cloud

- 8 ядер ~ 0.2 $/час

Работа в ограниченной памяти

scipy.sparse.csr_matrix – матрица в формате “по строкам”

- Массив непропущенных значений
- Массив индексов столбцов
- Массив промежутков, соответствующих строкам

Можно быстро выбирать подмножество строк

scipy.sparse.lil_matrix – вариант для конструирования

scipy.sparse.csc_matrix – матрица в формате “по столбцам”

Другие решения: ash & beluga (1 место)

Комбинации столбцов от станций S0-S11 и S12-S23

- Две группы станций имеют схожие структуры
- Можно попытаться комбинировать соответствующие столбцы

Кодирование числовых фич количеством примеров

Нормирование числовых фич по каждой неделе

Среднее расстояние во времени до ближайших N сбоев

Кол-во деталей на станции в заданном диапазоне времени

Другие решения: ash & beluga (1 место)

Downsampling строк без дубликатов

Обучение нескольких экземпляров модели в каждом фолде

“Leak-stratified k-fold” для кросс-валидации

- Разбиение на фолды с одинаковым количеством дубликатов

Другие решения:
"Data Property" Avengers (3 место)

Последовательности дубликатов по станциям

- Особенно помогло для S29 и S30

Target rolling mean при сортировке по мин/макс датам

- Особенно помогли большие окна

Bayesian mean для категориальных признаков

Кодирование пути прохождения через станции

Голосование 25 лучших сабмишнов

Другие решения: scnndl (4 место)

4 место менее чем за неделю и всего 12 сабмишнов!

- Большинство в топе сделало от 100 сабмишнов

Набор моделей первого уровня:

- По различным подмножествам станций:
- С/без числовых признаков
- С/без информации о дубликатах (в рамках подмножества)

Модели второго уровня обучаются только на дубликатах

Другие решения: scnndl (4 место)

Кросс-валидация

- На 1м уровне разбиение на фолды группами
- На 2м уровне обычный stratified k-fold
- Для выбора MCC - нижняя граница доверительного интевала в скользящем окне

Спасибо за внимание!