Решение конкурса BNP Paribas Cardif Claims Management

"Как успешно использовать скрипты для построения ансамбля"

42 место из 2947 в общем рейтинге

1 место среди студентов кафедры

Каюмов Эмиль

ММП ВМК МГУ

Семинар «Машинное обучение»

22 апреля 2016

Задача

- Задача: предсказать вероятность того, что заявку на страхование можно удовлетворить быстро без дополнительного рассмотрения.
- Анонимизированные данные, несбалансированные классы (3:1), 114000 объектов и 131 признак, много пропущенных значений.
- Метрика: LogLoss.

Фильтрация и наивный байес

1. Удалим сильно скоррелированные и малозначимые признаки.
2. Для категориальных признаков добавим признак-предсказание метода наивного байеса (кроме v22).
3. Пропуски заменим специальным значением.

Получили 112 признаков.

Можно было набрать public LB ≈ 0.452.

One-hot encoding

1. Слишком много уникальных значений признака v22 – объединим значения, встречающиеся менее 50 раз.
2. Сделаем one-hot encoding для категориальных признаков.
3. Удалим все прочие признаки.

Получили около 600 признаков.

На кросс-валидации можно было достигнуть ≈ 0.491.

Генерирование признаков линейной регрессией

1. Удалим скоррелированные признаки.
2. Объединяем признаки в подмножества до 3 признаков (наилучшие по качеству предсказания) и добавим предсказание линейной регрессии новым признаков.
3. Преобразуем v22 посимвольно.

Получили 137 признаков.

Можно было набрать public LB ≈ 0.451.

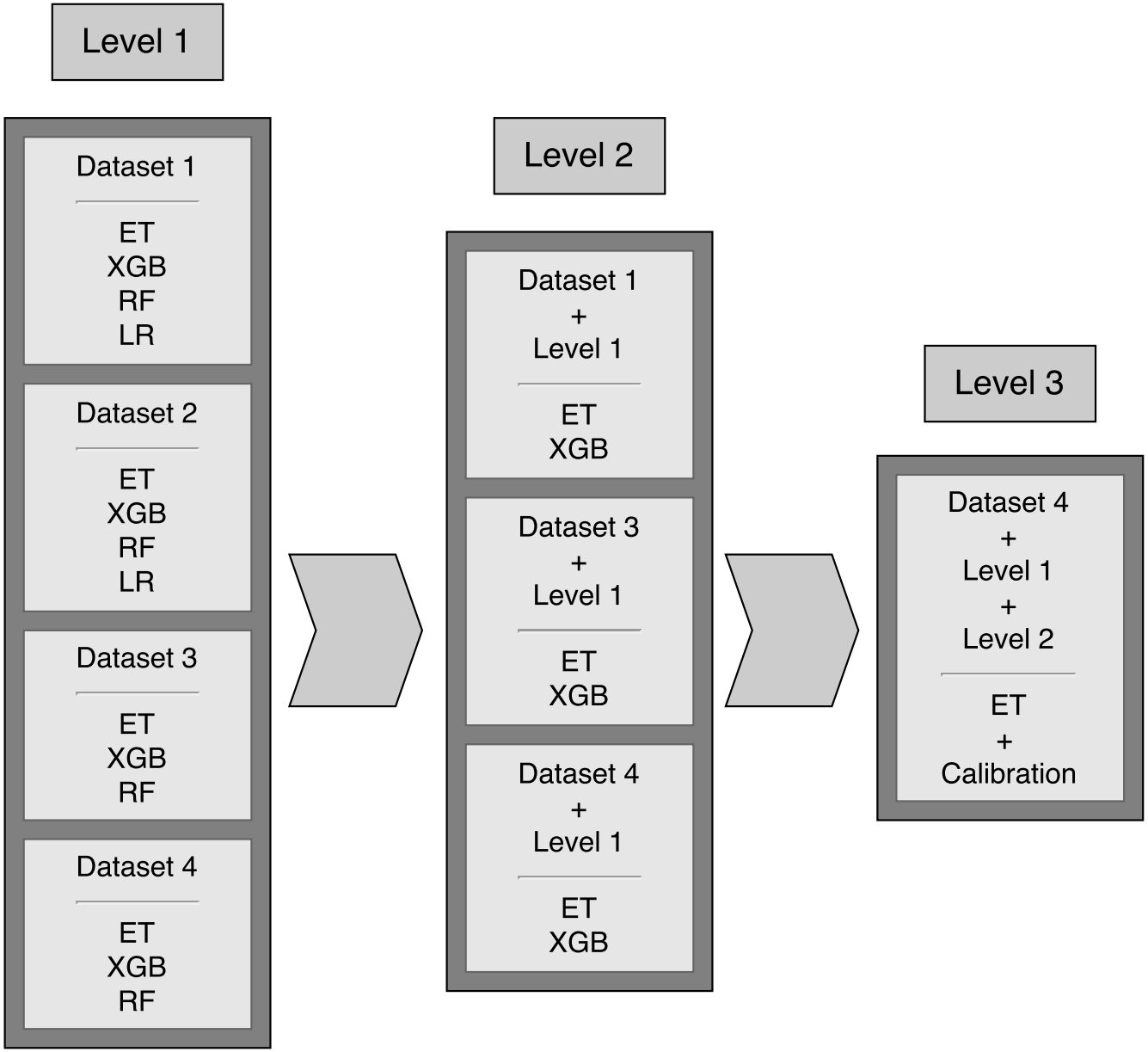
Ещё один датасет

1. Удаляем скоррелированные и малозначимые признаки.
2. Удаляем среди категориальных признаков те значения, которые не встречаются в тестовой выборке.
3. Добавляем признаки со средним значением и среднеквадратичным отклонением целевой переменной для каждого значения категориального признака, встречающегося более 50 раз.
4. Округляем действительные значения в шестом знаке и мастабируем.

Получили 46 признаков.

Можно было набрать public LB ≈ 0.455.

Архитектура



Результаты

Private LB: 0.43931.

- Обученный XGBoost на том же датасете давал результат, отличающийся лишь в пятом знаке после запятой.
- Блендинг на третьем уровне с моделями, обученными на третьем датасете, давали чуть худшие результаты.
- Калибрация вероятностей на модели без стэкинга давала прирост в четвёртом знаке после запятой.
- Первый уровень стэкинга давал прирост на 0.01, второй уровень вносил лишь небольшое улучшение результата.