Решение одного вечера

Competition 2, Yandex SHAD, Spring (1 место среди студентов ММП ВМК МГУ)

Каюмов Эмиль

ММП ВМК МГУ

Семинар «Машинное обучение»

22 апрель 2016

Задача

- Задача: определите наличие осложнений у пациента.
- Известны медицинские показатели и генетические данные о пациентах.
- Много признаков, мало объектов, несбалансированные классы.
- Метрика: LogLoss.

Уменьшение количества признаков

Мотивация: 1330 признаков на 4099 объектов?

- Обучим без подбора параметра случайный лес.
- Извлечём важность признаков.
- Отсечём по прогу наименее важные признаки.

Результат: 341 признак.

Пропущенные значения лучше заменять с помощью -1, чем наиболее частым значением.

Финальная модель

Используем случайный лес.

Private LB: 0.21491 (13 место).

Лучший незасчитанный сабмит

- Сделаем one-hot кодирование для категориальных признаков.
- Отбросим наименее важные признаки аналогично предыдущему разу.

Этим можно было добиться Private LB: 0.21439 (8 место).