Решение конкурса

Competition 2, Yandex SHAD, Spring

Амир Мирас Сабыргалиулы

ВМК МГУ

Курс «Машинное обучение, семинары» для 317-ой группы

Условие задачи

- По известным медицинским показателям, а так же генетическим данным о пациенте необходимо определить его состояние.
- Для каждого пациента задано описание в виде 1330 признаков:
  - Столбцы с V2 по V331 соответствуют различным медицинским показателям.
  - Столбцы с V332 по V1331 соответствуют генетическими данными пациента.
- Каждый признак принадлежит одному из трех типов:
  - Numeric - числовые признаки;
  - Category - порядок значений нам не важен;
  - Ordered Category - порядок значений важен;
- Качество решения оценивается с помощью функционала Logarithmic Loss

Предобработка данных

- Добавим индикаторы пропущенных значений
- LabelEncoder над медицинскими категориальными признаками
- Заполним пропущенные значения:
  - Вещественные признаки - средним
  - Категориальные признаки - самым часто встречающим
- Избавимся от генетических признаков, для которых порядок значений нам не важен
- Стандартизуем вещественные признаки

Скользящий контроль

- StratifiedKFold(n_folds=5)

Модель

- ExtraTreesClassifier(n_estimators=750, max_depth=8) над всеми признаки:
  - Результат на кросс валидации: **0.2156**
- XGBClassifier(max_depth=3, n_estimators=45) над всеми признаки, где медицинские признаки получены через LabelEncoder():
  - Результат на кросс валидации: **0.21847**
- Blending: $0.3 * \text{xgb} + 0.7 * \text{ext}$
  - Результат на кросс валидации: **0.21463**
  - Результат на public leaderboard: **0.23389**
  - Результат на private leaderboard: **0.21541**