ОТЧЕТ

Data Science problem

Автор
Черепахин Иван
icherepaxin@bk.ru

1 Постановка задачи

При заданных датасетах построить прогнозирующую модель. Указать в конечном файле вероятности принадлежности.

2 Этапы решения

При первоначальном анализе был сделан вывод, что задача является $multilabel\ classification$. Для ее решения я выбрал три модели: *kNN, decision tree, MLP*. Свой выбор могу пояснить тем, что данные модели хорошо работают с нашей задачей и также(что очень удобно), есть метод $predict\_proba$.
Далее выполнил обычную процедуру обработки и трансформации данных: избавление от NaN, поиск дубликатов, создании дамми переменных. В процессе было выявлено несколько закономерностей(которые логично было бы отнести к случайности). С пропусками я решил бороться просто их избавление(в будущем можно над этим поработать).
Переходя к моделям машинного обучения, я решил проверить на тренировочной выборке предсказательную способность. Я написал кастомную метрику, с помощью которой осуществлялась проверка задания. Она написана не оптимально и это требует доработки. Далее я написал примитивные(тк не имею больших вычислительных мощностей) пайплайны, которые тюнили гиперпараметры для каждой модели и провалидировал кастомной метрикой. Сетка для гиперпараметров выбиралась исходя из времени обучения и качеством метрики.

3 Итог

После проверки обучающей способности на тренировочном датасете был выбран MLP. Его же я применил уже на все выборке. Считаю, моя реализация не является конечной, тк не были до конца оптимизированы некоторые методы обработки данных. Также не была рассмотрена $calibrationcurves$, не была проведена проверка на других моделях деревьев(ансамблей).