Цифровой прорыв

Ярославская область

Ерохин А.И.

Решение (обработка данных)

1. Предобработка:
    1. Маппинг колонок для получения упорядоченных столбцов;
    2. Добавление признаков времени (время подъема и отхода ко сну, продолжительность сна);
    3. Замена редких численных значений на nan.

Решение (алгоритмы и подход)

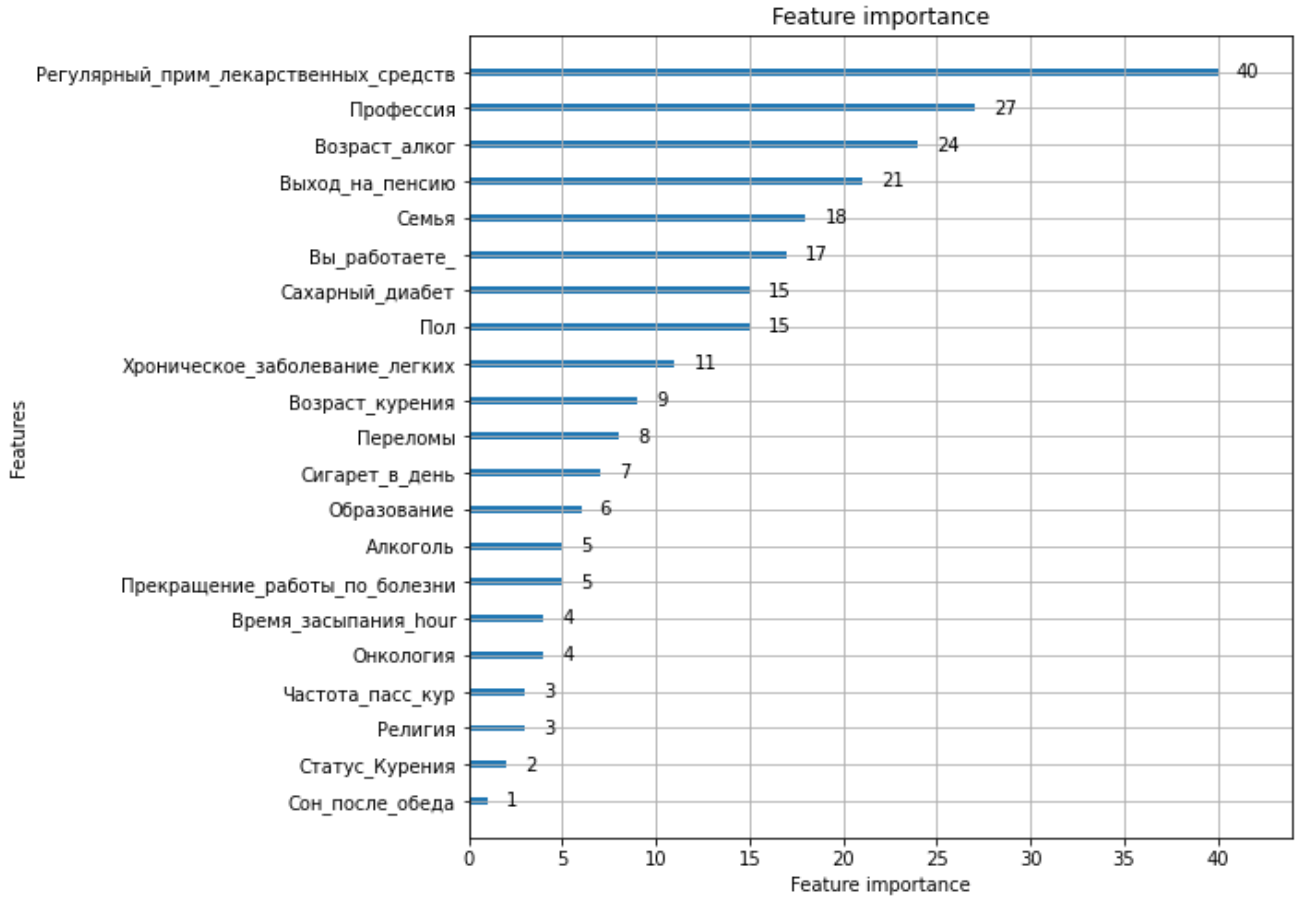
1. Алгоритмы и подход:
    1. Задача разделена на 2 – получение оптимальных параметров lightgbm и получение оптимальных порогов отсечения (при каких значениях округлять к 0 или 1);
    2. Параметры оптимизируются посредством optuna: для первой задачи – по roc-auc (т.к. это неплохая аппроксимация для целевой метрики), для второй – по целевой метрике задачи;
    3. Учитывая целевую метрику, один из классов будет важнее, чем другой в каждой из задач. Т.к. берется полусумма recall обоих классов – то цена ошибки не идентична;
    4. Для большей устойчивости при обучении используется RepeatedStratifiedKFold, то есть несколько стратифицированных разбиений;
    5. Предсказания моделей с шага 1 для каждого из Fold’ов усредняются при оптимизации на шаге 2 и предсказании на тесте.

Результаты

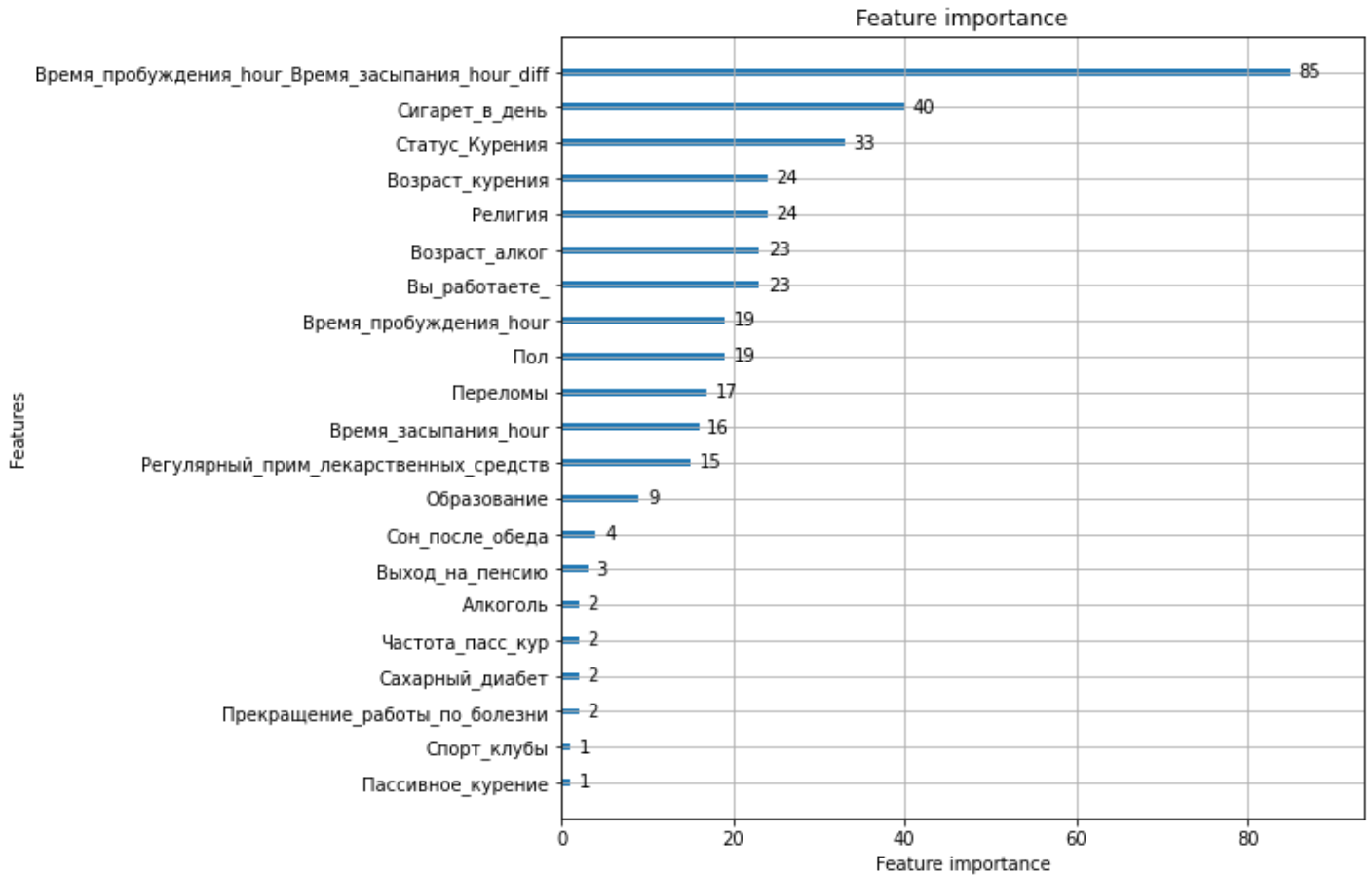
- Валидация:
  - Средняя целевая метрика: 0.653 (std 0.06);
- Лидерборд публичный:
  - Целевая метрика 0.664612

Учитывая специфику задачи, есть немалая вероятность т.н. shake-up, то есть перераспределения мест на приватной части лидерборда (из-за важности сложных в предсказании классов и последующего усреднения по всем 5 задачам классификации)

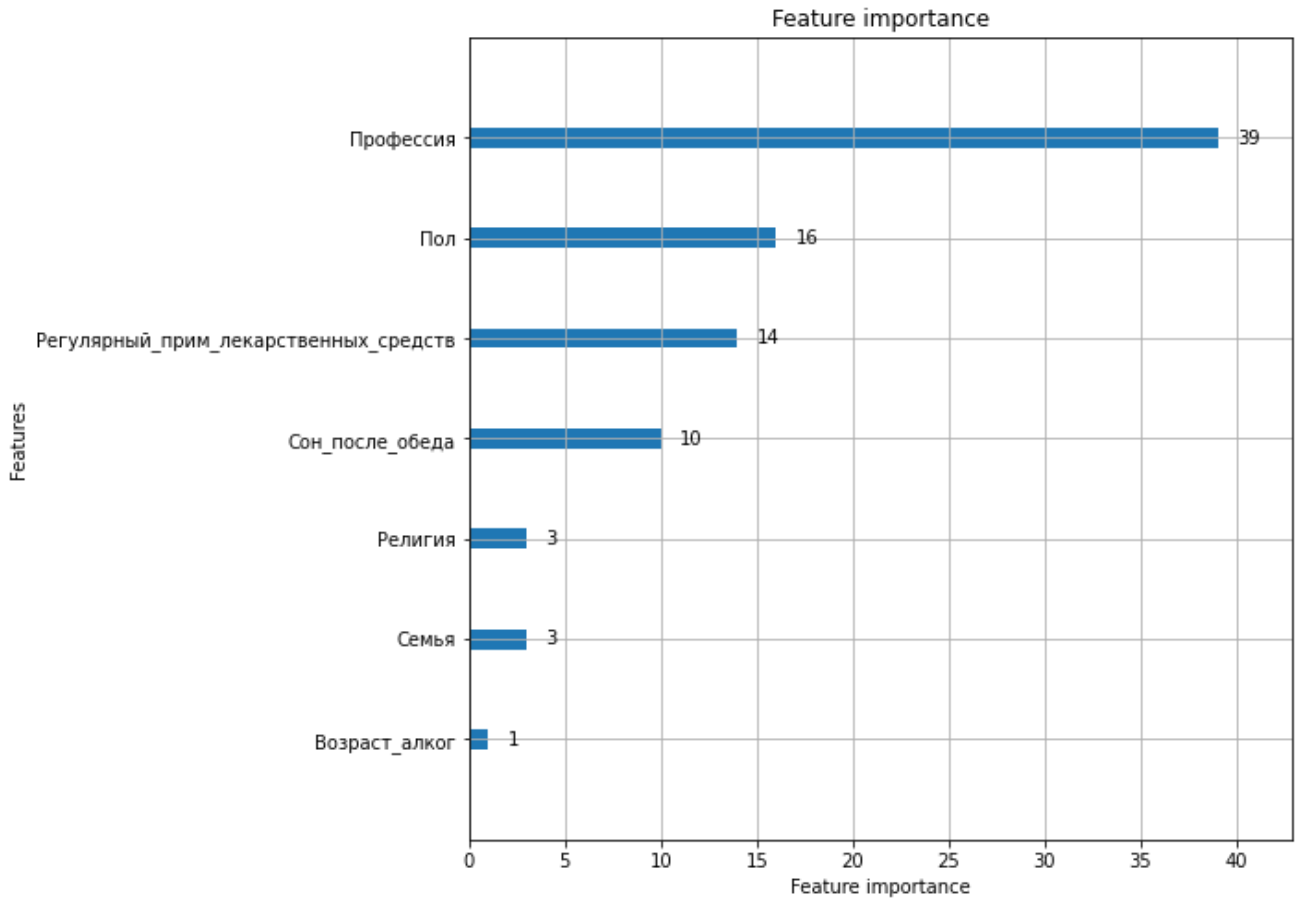
Важность признаков (класс Артериальная гипертензия)



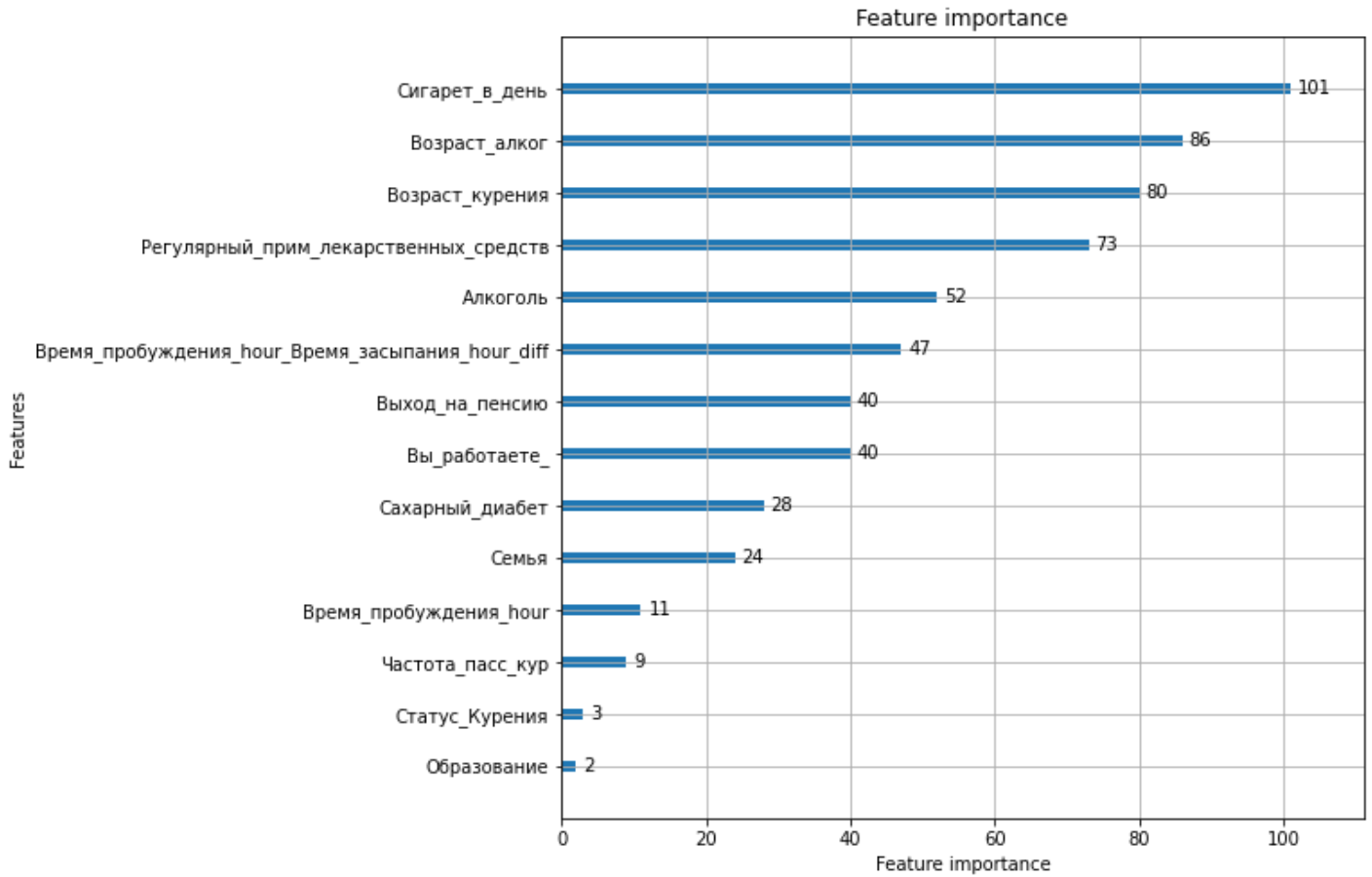
Важность признаков (класс ОНМК)



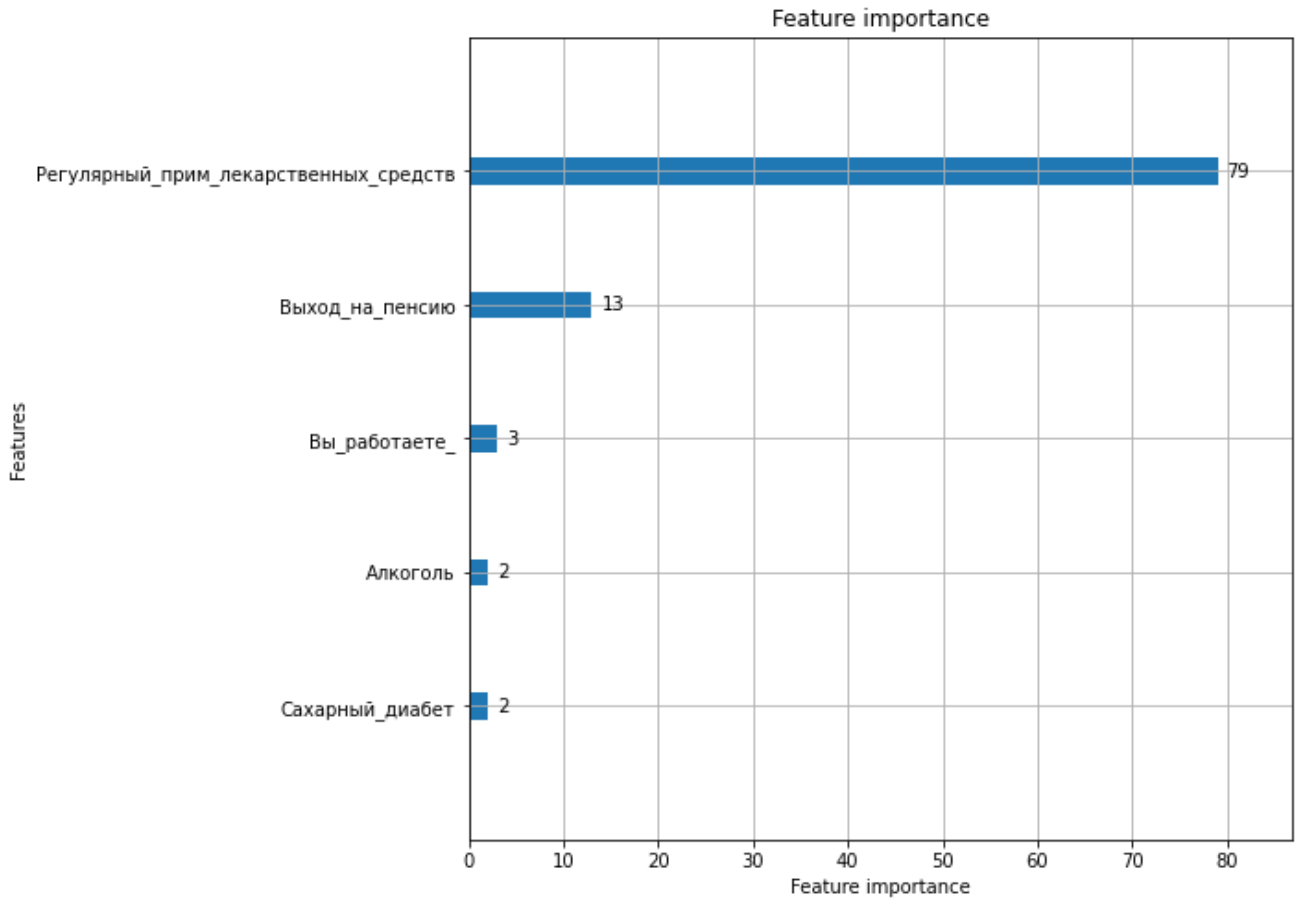
Важность признаков (класс Прочие заболевания сердца)



Важность признаков (класс Сердечная недостаточность)



Важность признаков (класс Стенокардия ИБС инфаркт миокарда)



Выводы

1. Целевая метрика кажется *не совсем корректной* в данном случае. Более корректным был бы выбор, основанный на важности соотношения ложных срабатываний и ложных пропусков;
2. Требуется *пересмотр процесса сбора данных*. Указанные признаки не всегда могут быть получены с адекватным уровнем качества (например, религия). Некоторые признаки могут иметь обратную связь с целевой переменной (например, прием лекарств может быть следствием наличия у людей диагноза и лечения, то есть причинно-следственная связь может быть обратной);
3. Требуется *уточнение процесса использования модели*, т.к. такого рода решения несут в себе некоторые морально-этические вопросы, которые должны решаться этическим комитетом.