Задача «Прогнозирование риска развития сердечно-сосудистого заболевания пациента»

Введение

Более 80% смертельных исходов в Российской Федерации связано с заболеваниями сердечно-сосудистой системы. Клинические методы диагностики не позволяют определять предикторы этих проблем без посещения медицинской организации и взятия анализов. В этой связи особую актуальность приобретает разработка интеллектуальной системы прогнозирования риска развития заболеваний с использованием исторических данных, позволяющей оценить риски критических нарушений в условиях повседневной жизни, без визита к врачу.

Распространенность болезней системы кровообращения преимущественно зависит от причин, главным образом, связанных с особенностями образа жизни, профилактическое воздействие на который может замедлить развитие заболеваний как до, так и после появления клинических симптомов. К настоящему времени выделяют более 300 факторов риска, однако, в первую очередь, сегодня используют семь факторов, вносящих основной вклад в преждевременную смертность: артериальная гипертензия, гиперхолестеринемия, курение, недостаточное потребление фруктов и овощей, избыточная масса тела, избыточное потребление алкоголя, гиподинамия. Это, так называемые, традиционные факторы. Если же традиционным факторам уделяется огромное внимание, то нетрадиционным посвящены всего лишь отдельные исследования.

Условие задачи

Задача — разработать модель машинного обучения, цель которой — предсказание наличия сердечно-сосудистых заболеваний. Данные содержат опрос реальных пациентов и их диагнозы, включая такие непрямые показатели, как образование, этнос, вид работы и многие другие. В рамках чемпионата вам необходимо классифицировать наличие/отсутствие у пациента следующих заболеваний:

- Артериальная гипертензия;
- Острое нарушение мозгового кровообращения;

- Стенокардия, ИБС, инфаркт миокарда;
- Сердечная недостаточность.

Описание входных значений

- train.csv — содержит в себе 1100 столбцов для обучения модели;
- test.csv — файл с пациентами, для которых необходимо сделать предсказание;
- sample_submission.csv — образец файла для отправки.

На что стоит обратить внимание

Данные содержат в себе множество непрямых показателей, с разной степенью вероятностью влияющих на наличие заболеваний пациентов. Возможно, от части столбцов имеет смысл отказаться в силу упрощения модели.

Метрика

В качестве метрики выступает Recall.

$$Recall = \frac{TP}{TP + FN}$$

Так как конечная задача преполагает мультиклассовое предсказание, то для каждой категории будет рассчитана своя метрика, которая будет усреднена. Такой полученный усредненный Recall для всего тренировочного датасета и будет является оценкой точности решения участника.