Проблема малых данных стоит параллельно теме больших данных, и даже фокус смещается со второй на первую. Такие задачи появляются часто, когда нужно выжить максимум из существующих данных, и уметь прогнозировать при принятии решения. Примеры, когда возникают такие задачи - дороговизна сбора данных, много времени или иных ресурсов, или размер данных ограничен на правовом уровне. И при работе с такими наборами обучающих сэмплов в таких обстоятельствах могут приводить к проблемам переобучения, понижения качества из-за шума, аномальных сэмплов, разреженности данных из-за пропусков. В моем отчете будет рассмотрена медицинская задача, где встретилась описываемая проблема, способы и результаты, полученные на текущий момент и возможные пути дальнейшей работы.

Передо мной была поставлена задача медиками центра Алмазова, изучающими эффективность препаратов по лечению ожирения, создать такую систему принятия решений, которая бы прогнозировала эффект от применения препаратов Сибутрамин и Саксенда и работала в условиях ограниченных данных. На разработку набора данных о пациентах, где были включены анализы крови, результаты опросников, сложные препараты на гормоны, были потрачены значительные бюджетные средства. Проводилось лечение пациентов и оценка параметров жизнедеятельности в течении от 3 и 6. Более того не все пациенты смогли пройти лечение весь курс 6 месяцев, что также снизило количество обучающих образцов.

Какие этапы включал в себя процесс создания такой системы - во-первых это отбор параметров за счет статистических тестов и постепенного удаления параметров и оценки эффективности модели. Далее происходило обучение моделей прогнозирования (сначала регрессии, но точность таких моделей была очень низкая, затем я пришел к классификации), оценка метрик классификации. Для медиков важно знать точный промежуток процента потери веса - была произведена многоклассовая классификация. Результаты представлены в таблице - для препарата Саксенда удалось создать модель с точностью 60% что устраняет неопределенность при принятии решения, а для препарата Сибутрамин такого не получилось. Далее был реализован метод объяснения моделей методом чисел Шапли и эти модели были собраны в приложение. Веб приложение сейчас в стадии разработки.

Если вернуться к результатам модели, то видно, что еще есть потенциал к улучшению эффективности модели, но он требует более продвинутых методов. В этом состояла цель моей исследовательской работы - поиск метода устранения проблемы малых данных при обучении модели машинного обучения в системе. Для этого я проанализировал существующие методы, применяемые в научных статьях.

Рассмотрим нейросетевые методы – первое возможное решение проблемы - сжатие данных с максимальным сохранением закономерностей. Я хочу рассмотреть ансамбль автоэнкодеров, который разделяет пространство признаков и создает такое представление данных, устраняя переобучение и проблемы малых данных. Недостаток - настройка разбиения, ухудшение интерпретируемости модели. Далее возможным решением проблемы может являться генерация табличных данных, которая может вывести задачу из ограничений малых данных - недостаток, не всегда можно обучить хорошую модель, особенно на малых данных, если в итоговых данных закономерности не соблюдены.

Следующее решение - методы фильтрации. В ситуации малых данных важно чтобы каждый сэмпл обучающей выборки улучшал модель и не нарушал процесс обучения. И такую проблему поиска аномалий можно решать методами кластеризации, где более удаленные сэмплы от центра кластера удаляются в процессе обучения. Усложненный вариант — это генерация сэмплов с последующей фильтрацией. Этот метод более перспективный, так как включает в себя методы генерации и фильтрации одновременно, и может значительно улучшать модели в медицинской сфере. И последнее решение — это модель, не входящая в множество ML - Баесовская модель оценки вероятностей, где математическими методами происходит дискредитизация данных, составляются совместные распределения результатов классификации и разбиения параметров. Это самый сложный метод из представленных, и будет рассматриваться в последнюю очередь. Цель на преддипломную практику оценка эффективности перечисленных методов в задаче улучшения качества модели и внедрение такого метода в итоговое приложение.

Вывод: на данный момент разработки мне удалось обучить модели, которые прогнозируют отклик на лечение в виде моделей классификации. Классическими методами удалось добиться хорошего качества прогнозирования для препарата Саксенда, а для Сибутрамин качество хуже - 47%. Необходимо дальнейшее улучшение системы (решение проблемы малых данных), цель моей исследовательской работы - поиск такого метода, текущий результат - обзор нейросетевых, генеративные, фильтрующие, баесовских методов, которые могут улучшить существующие результаты и сделать рекомендацию более качественной. В итоге планируется внедрение найденного метода, оформление приложения для медиков и патента на готовое приложение.