МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В.ЛОМОНОСОВА
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ

УТВЕРЖДАЮ
Декан факультета ВМК МГУ,
академик РАН

И.А. Соколов

« » _____ 2024 г.

ОТЧЕТ

О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ ПО ДОГОВОРУ №6.2.18 (ГОСБЮДЖЕТ, № ЦИТИС АААА-А18-118011590152-8) ИССЛЕДОВАНИЕ, РАЗРАБОТКА И ПРИМЕНЕНИЕ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ ПОСТРОЕНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ ПРОГРАММНЫХ СИСТЕМ

Руководитель проекта,
д.ф.-м.н., профессор, заведующий кафедрой И.В. Машечкин

Москва 2024

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель проекта, д.ф.-м.н., профессор, заведующий кафедрой	И.В. Машечкин
Исполнители темы	
к.ф.-м.н., доцент	М.И. Петровский
д.т.н., профессор	А.П. Рыжов
м.н.с	И.С. Попов
к.ф.-м.н., доцент	М.А. Казачук
математик	Ю.А. Васильев
инженер	О.Е. Горохов
математик	И.С. Лазухин
инженер	С.В. Герасимов
к.ф.-м.н., математик	Д.В. Царев

РЕФЕРАТ

Отчет 77 с., 1 ч., 33 рис., 6 табл., 54 источника.

COVID-19, ПРОГНОЗИРОВАНИЕ, ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, МАШИННОЕ ОБУЧЕНИЕ, ГРАДИЕНТНЫЙ БУСТИНГ, НЕЙРОННЫЕ СЕТИ, КОМПЬЮТЕРНАЯ ТОМОГРАФИЯ.

В отчете содержится информация о проведенных работах в части доработки функционала информационной системы - сервис «Калькулятор КТ», предназначенного для получения экспресс-оценки изменений легочной ткани при COVID-19 без применения компьютерной томографии органов грудной клетки на основе физикальных и лабораторных анализов пациента, созданного на предыдущих этапах исследований:

- Формирование проверочных выборок и проведение оценки и доработки моделей-кандидатов;
- Доработка моделей машинного обучения и выбор лучшей модели для программного сервиса.

в рамках задачи исследования и разработки методов искусственного интеллекта и анализа больших данных в сфере здравоохранения.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями.

ТЗ	Техническое задание
СПО	Системное программное обеспечение
ПАО	Программно-аппаратное оборудование
ИС	Информационная система
UI	User Interface (пользовательский интерфейс)
КТ	Компьютерная томография
ПЦР	Полимеразная цепная реакция – метод молекулярной биологии, позволяющий добиться значительного увеличения малых концентраций определённых фрагментов нуклеиновой кислоты (ДНК или РНК) в биологическом материале (пробе)
ИФА	Иммуноферментный анализ – лабораторный иммунологический метод качественного или количественного определения различных низкомолекулярных соединений, макромолекул, вирусов и пр., в основе которого лежит специфическая реакция антиген-антитело
SpO2	Сатурация – насыщение крови кислородом
RDW	Ширина распределения эритроцитов
HGB	Значение Гемоглобина
MPV	Средний объем тромбоцитов в крови
RBC	Количество эритроцитов
PLT	Количество тромбоцитов
PDW	Ширина распределения тромбоцитов по объему
WBC	Количество лейкоцитов
CRB	Уровень С-реактивного белка
АЛТ	Значение аланинаминотрансферазы
АСТ	Значение аспартатаминотрансферазы
URO	Уробилиноген
ЧДД	Частота дыхательных движений

ОАК	Общий анализ крови
IGG	Иммуноглобулины класса G
IGM	Иммуноглобулины класса M
ЛИ	Лабораторные исследования
NN	Neural Network (нейронная сеть)
RF	Random Forest (случайный лес)

ВВЕДЕНИЕ

Всемирная организация здравоохранения 11 марта 2020 года объявила пандемию по заболеванию COVID-19, вызываемому вирусом SARS-CoV-2. Пандемия резко ускорила внедрение цифровых сервисов в работу московского здравоохранения. Начиная с марта 2020 года по декабрь 2020 года московские врачи вылечили свыше 500 тысяч человек – полмиллиона электронных историй болезни и выздоровления, в которых накоплен огромный объем данных. Вследствие этого актуальным является исследование возможности применения технологий искусственного интеллекта и машинного обучения для создания полезных инноваций для спасения жизни людей.

Одним из ключевых результатов предыдущих исследований в рамках данной научно-исследовательской работы было создание КТ-калькулятора – технологии оценки степени изменения легочной ткани при COVID-19 в экспресс режиме без использования инструментальных методов исследования, позволяющего на основе анализов крови, сатурации и общей клинической картины пациентов спрогнозировать вероятность легкого (КТ 0-1), среднего (КТ 2) или тяжелого (КТ 3-4) течения пневмонии и принять решение о дальнейшей тактике лечения. Данный сервис позволяет сократить количество исследований для пациента и уменьшить облучение. Также, он сокращает нагрузки на реальное оборудование – на КТ-центры и применим в тех регионах, в которых доступ к компьютерному томографу ограничен, либо этого оборудования нет. Точность результата составляет порядка 90%.

Поскольку количество электронных историй болезни и выздоровления ежедневно пополняется, актуальным является использование новых данных COVID-больных для доработки функционала работы КТ-калькулятора с целью дальнейшего повышения качества его работы.

В рамках настоящего этапа данной работы исследованы и разработаны методы интеллектуального анализа данных для решения следующих актуальных задач:

- Формирование проверочных выборок и проведение оценки и доработки моделей-кандидатов на включение в программный сервис;
- Доработка моделей машинного обучения и выбор лучшей модели для программного сервиса.

1 Аннотация

В рамках данного этапа работ были проведены научно-исследовательские работы в области создания прогностических моделей течения и исхода заболевания у пациентов с COVID-19. На основе данных о клинических особенностях, факторах коморбидности, клинико-лабораторного анализа и других факторов, потенциально связанных с тяжестью течения заболевания и вероятностью смерти пациентов с COVID-19, был разработан комплекс моделей, построенных с использованием передовых методов машинного обучения и прикладного статистического анализа, для прогнозирования тяжести течения и исхода заболевания у пациентов, получающих лечение в амбулаторных и стационарных условиях.

Одним из ключевых результатов предыдущих исследований в рамках данной научно-исследовательской работы было создание КТ-калькулятора – способа оценки степени изменения легочной ткани при COVID-19 в экспресс режиме без использования инструментальных методов исследования, и позволяющего на основе анализов крови, сатурации и общей клинической картины спрогнозировать степень тяжести пневмонии пациентов.

Основной целью настоящего этапа исследований являлась доработка «КТ-калькулятора» с учетом дополнительных наборов данных.

Были решены все поставленные задачи:

- Формирование проверочных выборок и проведение оценки и доработки моделей-кандидатов на включение в сервис «Калькулятор КТ»;
- Доработка моделей машинного обучения и выбор лучшей модели для сервиса «Калькулятор КТ».

2Постановка задачи

1.1 Формирование проверочных выборок и проведение оценки и доработки моделей-кандидатов на включение в сервис «Калькулятор КТ», разработанный на предыдущих этапах исследований и дорабатываемый на текущем этапе с использованием следующих источников данных:

- Данные о госпитализированных пациентах, собранные в результате экспериментальной работы с сервисом «Калькулятор КТ» в ГБУЗ «ГКБ № 67 им. Л.А. Ворохобова ДЗМ»;
- Данные об амбулаторных пациентах, собранные в результате экспериментальной работы с сервисом «Калькулятор КТ»;
- Данные о пациентах из набора открытых данных изображений КТ легких при COVID-19 Технического университета Хуажонг [2].
- Данные ДИТ от 15.02.21 обо всех клинических и амбулаторных COVID-19 пациентах:
    - общие данные по выжившим и умершим пациентам (включая анамнез) с марта 2020 года до февраля 2021 года;
    - данные о проведенных клинических и биохимических анализах с марта 2020 года до февраля 2021 года;
    - данные из КТ-центров (степень тяжести, а также осмотровые признаки) с марта 2020 года до декабря 2020 года;
    - данные о проведенных тестах ПЦР и ИФА с марта 2020 года до декабря 2020 года;
    - данные о взятой сатурации с марта 2020 года до февраля 2021 года.

1.2 Доработка моделей машинного обучения и выбор лучшей модели для сервиса «Калькулятор КТ» на основе проверки следующих вариантов:

- Раздельное прогнозирование тяжести только по анализу крови с помощью нейросети и только по результатам осмотра с помощью ансамблей деревьев решений (алгоритм машинного обучения «случайный лес») для последующего объединения откликов моделей через ансамбль;

- Использование бустинг ансамбля деревьев решений lgbm (алгоритм машинного обучения «градиентный бустинг») для прогнозирования степени тяжести КТ;
- Использование ансамбля регуляризованных нейросетей для прогнозирования степени тяжести КТ;
- Дополнительная очистка и нормализация тренировочной выборки с удалением противоречий;
- Расширение признакового пространства, в том числе за счет включения информации о датах проведения анализов и осмотра;
- Взаимообратный учет показателя сатурации в оценке степени тяжести с пересчетом по баллам шкалы NEWS;
- Алгоритмы для автоматизированного тьюнинга и регуляризации моделей;
- Исследование различных видов калибровок результата и выбора порогов;
- Методы непротиворечивого сведения откликов двух независимых бинарных моделей, прогнозирующих отдельно вероятность тяжелого поражения (КТ34) и легкого поражения (КТ01) к многоклассовой вероятностной модели (КТ01, КТ2, КТ34) и к единой шкале оценки тяжести;
- Разработка окружения и скриптов для автоматизированного тестирования сервиса «Калькулятор КТ» через интерфейс REST API;
- Разработка средств «объясняющей» визуализации результатов прогноза на основе методов LIME и kernel SHAP, применение и анализ результатов этих методов к ошибкам прогноза.

3 Исследование и построение решения

3.1 Формирование проверочных выборок и проведение оценки и доработки моделей-кандидатов

3.1.1 Описание наборов данных

3.1.1.1 Данные о госпитализированных пациентах, собранные в результате экспериментальной работы с сервисом «Калькулятор КТ» в ГБУЗ «ГКБ № 67 им. Л.А. Ворохобова ДЗМ»

- Объем данных: 20 Кб
- Количество наблюдений: 95
- Содержит следующие столбцы:
  - **uuid** – уникальный идентификатор
  - **Анамнез: Пол, Возраст, хронические заболевания (ишемическая болезнь сердца, ожирение, сахарный диабет);**
  - **Осмотровые признаки: степень тяжести, сатурация, ЧДД, наличие одышки, слабость, заложенность, температура тела, наличие кашля, тип кашля;**
  - **Лабораторные анализы: PCR, WBC, PDV, MON, GRA, LYM, PLT, HGB, RBC, MPV, HCT, RDW**

Данный набор данных может быть использован в виде тестовой выборки для оценки качества моделей, так как для обучения модели наблюдений слишком мало и выборка имеет явный дисбаланс в сторону степени тяжести КТ 1.

3.1.1.2 Данные об амбулаторных пациентах, собранные в результате экспериментальной работы с сервисом «Калькулятор КТ»

Разработанный сервис «Калькулятор КТ» содержит инструмент логирования всех поступивших запросов. Каждый поступивший в сервис запрос записывается в файл формата «getForecast[YYYY-MM-DD]», где YYYY-MM-DD – дата поступления запроса (например, «getForecast[2021-03-09].txt»).

Каждое логируемое событие включает:

- Контекстные признаки: время запроса, ip адрес источника, уникальный идентификатор uuid (генерируется при поступлении запроса и успешного применения модели прогнозирования), тип прогноза;
- Контентные признаки: входные признаки и результат прогноза классификаторов КТ 01, КТ 34.

Уникальный идентификатор также возвращается пользователю сервиса «Калькулятор КТ». Так как на данный момент в сервис не встроена система оценки корректности прогноза, обратная связь от пользователя может быть получена по идентификатору uuid.

При получении ошибочного результата, пользователь может сообщить разработчику свой uuid, после чего разработчиком будет проводиться анализ логированных данных о совершенном запросе (в частности, проверка на противоречивость). Если обращение корректно, событие с ошибкой может быть внесено в тестовый набор, на основе которого будет проводиться тестирование последующих моделей.

В дальнейшем, возможно организовать дообучение модели на поступающих данных. Однако, в данном случае, необходимо вводить аппарат получения разметки событий и встраивать данный аппарат в пользовательский интерфейс веб-формы Калькулятора КТ.

3.1.1.3 Данные о пациентах из набора открытых данных изображений КТ легких при COVID-19 Технического университета Хуажонг [2]

Данный набор данных содержит информацию об анализах крови, клинических особенностях и изображениях компьютерной томографии грудной клетки больных COVID-19 пациентов, находящихся в больницах Китая. Поскольку в данном наборе данных содержатся только сами изображения КТ, врачами ГКБ №67 была проведена ручная разметка соответствующих данных. Результирующий набор данных содержит следующие признаки:

- Температура тела;
- Пол;
- Возраст;
- ПЦР;
- Абсолютное число лимфоцитов;
- Ширина распределения эритроцитов (RDW);
- Значение гематокрита;
- Наличие сахарного диабета;

- Наличие ожирения;
- Абсолютное количество моноцитов;
- Значение Гемоглобина (HGB);
- Средний объем тромбоцитов в крови (MPV);
- Количество тромбоцитов (PLT);
- Ширина распределения тромбоцитов по объему (PDW);
- Количество лейкоцитов (WBC);
- Количество эритроцитов (RBC);
- Уровень С-реактивного белка (CRB);
- Значение креатинина.

Размеченная выборка содержит данные 185 пользователей. Данный набор данных может быть использован в виде тестовой выборки для оценки качества моделей, так как для обучения модели наблюдений слишком мало и выборка имеет явный дисбаланс в сторону степени тяжести КТ 1.

3.1.1.4 Данные ДИТ от 15.02.21 обо всех клинических и амбулаторных COVID-19 пациентах

Набор данных представляет собой выгрузку из системы ЕМИАС с 03.2020 до 02.2021.

Выгрузка по амбулаторным пациентам имеет следующую файловую структуру:

- Данные из КТ-центров
  - Объем данных: 92 Мб
  - Количество наблюдений: 303’628
  - Содержит следующие столбцы:
    - **uuid** – уникальный идентификатор
    - **event_start_time –** время проведения КТ
    - **temperatura_tela_time –** время измерения температуры
    - **temperatura_tela_value –** значение температуры
    - **osnovnoy_diagnoz, soputstvujuschij_diagnoz –** диагнозы пациента в виде кода МКБ-10
    - **drugie_soputstvujuschie_diagnozy{1-7} –** другие сопутствующие диагнозы в виде кода МКБ-10
    - **has_kashel, kashel_type –** наличие и тип кашля (сухой, с мокротой)
    - Бинарные признаки одышки, заложенности, слабости

        - **chdd –** частота дыхательных движений
        - **osmotr_tyazhest –** тяжесть пациента при осмотре (без симптомов, легкая, средняя, тяжелая)
        - **KT_nalichie_pnevmonii –** бинарный признак наличия пневмонии
        - **KT_stepen_tjazhesti –** степень тяжести КТ (нулевая, легкая, средне-тяжелая, тяжелая, критическая)
        - **resultat_KT –** результат КТ (КТ-0, КТ-1, КТ-2, КТ-3, КТ-4)
- Данные по проведенным амбулаторным анализам
    - Объем данных: 13,5 Гб
    - Количество наблюдений: 43’348’273
    - Содержит следующие столбцы:
        - **uuid** – уникальный идентификатор
        - **event_start_time –** время проведения анализа
        - **nazvanie_issledovanija –** название исследования
        - **test_time –** время сбора анализа
        - **nazvanie_testa –** название теста
        - **znachenie_rezultata_ed_izm –** единицы измерения результата
        - **znachenie_rezultata –** результат теста
        - **referensnye_znachenija –** референсные значения
- Данные по проведенным тестам ПЦР и ИФА
    - Объем данных: 2.4 Гб
    - Количество наблюдений: 4’466’407, в частности:
        - 3’346’444 наблюдения по ПЦР
        - 1’103’446 наблюдения по ИФА
    - Содержит следующие столбцы:
        - **uuid** – уникальный идентификатор
        - **birth_dt –** дата рождения
        - **gender –** пол
        - **pcr_ifa –** тип теста: ПЦР или ИФА
        - **mu_name –** наименование пункта сбора тестов
        - **department_name –** наименование филиалов
        - **dis_date –** дата поступления
        - **get_date_at, get_time_at, send_date_at, send_time_at –** даты получения и выдачи результатов теста

        - **mkb10_name, mkb10_code –** название и код выявляемого диагноза
        - **samples_type –** тип сбора анализа
        - **samples_result –** результат анализа
        - **laboratory_name –** наименование лаборатории
- Данные по взятой Сатурации
    - Объем данных: 46 Мб
    - Количество наблюдений 661’353
    - Содержит следующие столбцы:
        - **uuid** – уникальный идентификатор
        - **event_start_time**
        - **saturation**
- Общие данные по выжившим пациентам
    - Объем данных: 320 Мб
    - Количество наблюдений: 880’352
    - Содержит следующие столбцы:
        - **uuid** – уникальный идентификатор
        - **Причина закрытия ЭС**
        - **Группа риска**
        - **Дата рождения**
        - **Пол**
        - **Дата взятия первого анализа**
        - **Дата получения результата первого анализа**
        - **Результат повторного анализа**
        - **Дата получения повторного анализа**
        - **Дата получения предпоследнего результата ПЦР**
        - **Результат предпоследнего анализа ПЦР**
        - **Результат последнего анализа ПЦР**
        - **Тяжесть заболевания**
        - **Тяжесть заболевания.1**
        - **Тяжесть заболевания.2**
        - **Тяжесть заболевания.3**
        - **Тяжесть заболевания.4**
        - **Откуда пришел в стационар**
        - **Дата госпитализации в стационар**

        - **ОРИТ**
        - **ИВЛ**
        - **ЭКМО**
        - **Дата размещения в обсерваторе**
        - **Характеристика статуса в поликлинике**
- Общие данные по умершим пациентам
    - Объем данных: 8 Мб
    - Количество наблюдений: 48’415
    - Содержит следующие столбцы:
        - **uuid** – уникальный идентификатор
        - **Дата смерти**
        - **Дата рождения**
        - **Пол**
        - **Причина смерти {А-Д}**
        - **Код смерти {А-Д}**
        - **РФС**
        - **Ковид**
        - **Основная причина смерти**
        - **Основной ковид**

Набор данных является целевым для обучения и тестирования прогнозных моделей. Основными факторами являются: размер набора данных (как до, так и после балансировки), большой спектр возможных предикторов и заполненность полей (с возможным дополнительным заполнением пропусков на основе неиспользуемых в модели предикторов). Формирование дальнейших описываемых выборок будет основываться на данном наборе. Наборы данных, описанные в пунктах 3.1.1.1, 3.1.1.3, используются для тестирования лучших прогнозных моделей (процесс тестирования описан в пункте 3.2.10). Набор данных, описанный в пункте 3.1.1.2, используется для корректной обработки обратной связи от пользователей Калькулятора КТ и построения статистических показателей.

3.1.2 Подготовка данных

На вход подаются данные медицинских анализов пациентов из набора ДИТ от 15.02.21 обо всех клинических и амбулаторных COVID-19 пациентах.

Необходимо подготовить эти данные перед построением математических моделей:

- *Очистить данные и привести значения показателей к общим шкалам и словарям;*
- *Найти, удалить и исправить артефакты, выбросы и противоречивые данные.*

На основе исходных наборов данных формируется набор данных для прогнозирования КТ на основе физикальных и клинических признаков.

1. Для каждого наблюдения из КТ центра производится поиск ближайших (окно – неделя) тестов пациентов:
    a. Тесты фильтруются по выделенным группам
    b. На основе uuid производится срез тестов пациента
    c. Тесты сортируются по близости даты к дате проведения КТ
    d. Выбирается ближайший тест
    e. Каждому тесту соотносится 3 признака: значение теста, референсные значения, дата взятия теста.
2. В качестве тестов пациентов взяты следующие группы анализов:
    a. Тесты общего клинического анализа крови
    b. Наиболее заполненные тесты биохимического анализа (определение аланинаминотрансферазы (АЛТ), определение альбумина, определение аспартатаминотрансферазы (АСТ), определение билирубина общего, определение билирубина прямого (конъюгированного) моноглюкоронида и диглюкоронида, определение калия общего, определение креатинина, определение лактатдегидрогеназы, определение мочевины, определение натрия общего, определение общего белка, определение хлора, определение щелочной фосфатазы, относительное количество нормобластов)
3. По всем выделенным тестам производится фильтрация исходного набора данных. Таким образом, производится поиск теста (а не анализа), возможно получение результатов теста от нескольких источников исследований.
   Например, Гематокрит встречается в нескольких исследованиях.
4. На основе словарей слияния производится объединение нескольких тестов (от разных исследований), в один признак теста, а размерность значений приводится к одной (наиболее часто встречающийся).
   Например, Нитриты имеют запись в двух форматах: «Нитриты» и «Нитриты (NIT)». Значения в форматах могут иметь разные размерности. Тогда итоговый признак «Нитриты» берет за основу формат с лучшей заполненностью и

обогащает его значениями из других форматов, приведенных к общей размерности признака.

5. Выделяются наиболее заполненные признаки среди близких тестов.

   Например, для тромбоцитов существует множество признаков после объединения тестов от разных исследований:

   - Общий объем тромбоцитов в крови (тромбокрит, PCT)
   - Количество тромбоцитов
   - Средний объем тромбоцитов в крови
   - Ширина распределения тромбоцитов по объему

   Так как данные признаки коррелируют между собой, в ходе формирования признакового пространства выбирается наиболее заполненный признак.

6. Сформированный набор дополняется тестами ПЦР, ИФА аналогичным образом (выделение для каждого пациента ближайших тестов с окном – неделя)
7. Сформированный набор дополняется Сатурацией аналогичным образом (выделение для каждого пациента ближайшей Сатурацией с окном – неделя)
8. Формируются признаки хронических болезней:
   a. Для каждого пациента формируется признак диагнозов – список всех кодов МКБ-10, сопоставленных данному пациенту.
   b. Выявляются хронические диагнозы по следующим множествам (болезнь – код МКБ-10):
      - ишемическая болезнь сердца [5]: I11 I20 I24 I25 I51
      - артериальная гипертензия [6]: I10 O10-13 G97 I27 K76 P29 I15
      - сахарный диабет [7]: G63 E10-14 H36 M14 G59 E23 N08 O24
      - хронические болезни легких [8]: J44, J45, J47, J42, J60-70
      - ожирение [9]: E66
   c. Признаки болезней бинаризуются
   d. Согласно сформированным признакам, хронические заболевания имеются у следующих пациентов:
      - ишемическая болезнь сердца: 19’120 пациента
      - артериальная гипертензия: 278 пациентов
      - сахарный диабет: 6’224 пациента
      - хронические болезни легких: 0 пациентов
      - ожирение: 1106 пациентов

3.1.2.1 Очистка и приведение значений показателей к общим шкалам и словарям, включая учет референсных значений

- Данные из КТ-центров
  - В поле **temperatura_tela_value** имеются значения температуры: 3.0, 3.6, 3.8.
  - В поле **chdd** имеются следующие значения ЧДД: 0, 1 и более 150.

    Решение: удаление данных наблюдений.
- Данные по проведенным амбулаторным анализам
  - Неунифицируемость единиц измерения (**znachenie_rezultata_ed_izm**).

    Встречаются как стандартные единицы измерения («фл», «10^12/л», «Ед/мл» и т.д), так и их различные вариации («fL», «усл.ед», «млн», «мг%»).

    Также, встречаются не интерпретируемые размерности:

    «/L», «/Ед», «1/поле зрения высокого увеличения», «%10^9л».

Решение: предложен следующий алгоритм обработки единиц измерения (далее е.и.) признака **N:**

1. Выявление наиболее частых е.и. данного признака
2. Выделение префиксов и постфиксов данной е.и.
3. Для остальных е.и. вычисляется расстояние до исходной е.и.
   a. Обработка префиксов и постфиксов «м, мк, н, мл, к, М»
   b. Обработка степенных значений: 10^9 и т.д.

- Неунифицируемость формата результатов (**znachenie_rezultata**)

  Помимо вещественных чисел, возможны и категориальные оценки выявления (обнаружено, не обнаружено), но данные категориальные оценки не имеют унифицируемого формата.

  Например, смысловое значение «не обнаружено» может иметь следующие вариации: «Не обнаружено», «не обнаружено», «не обнаружены», «нет», «Нет», «0 (Отрицательно)», «Отрицательно», «ОТРИЦАТЕЛЬНО», «не обнаружен», «-», «Отрицательный», «0 (Не обнаружено)», «Не обнаруж,», «не обнаружена», «Отсутствует», «не обн.».

  Данные вариации встречается не менее 5000 раз.

Решение: предложен следующий алгоритм обработки результатов:

1. Для категориальных переменных составлены словари значений, все возможные значения приведены к нижнему регистру и приводятся к единым категориям.

2. Для непрерывных признаков производится удаление незначащих символов, приведение к вещественному виду. В случае множественных значений, производится разбиение по разделителям и выбирается первое значение.
3. Если в ходе обработки возвращена ошибка – данное значение заменяется средним по обучающей выборке.

- Неунифицируемость формата референсных значений (**referensnye_znachenija**).

  В качестве референсных значений могут быть указаны следующие категории представлений:

  - Повторение результатов теста:
    «отрицательно», «не обнаружено», «не обнаружены\n\n» и т.д.
  - Интервалы и промежутки:
    «<10», «10-20», «3,5 - 6,1», «2.04 - 5.80», «меньше 2000»
  - Полная сводка возможных интервалов:
    «Мужчины: 0,0-15,0. Женщины: 0,0-20,0.», «отр.<9 пол.>11»,
    «близко к оптимальному уровню 2,6 - 3,3\nоптимальный уровень < 2,6\nпограничный уровень 3,3 - 4,1\nвысокий уровень 4,1 - 4,9\nочень высокий уровень > 4,9»
  - Смесь интервалов с размерностями:
    «<34 мкмоль/л»
  - Перечень категорий:
    «Кислая, слабокислая, нейтральная», «светло–желтый, желтый, соломенно-желтый»
  - Некорректные значения:
    «отрицательно, исслед. не проводилось», «оформленный»
  - И т.д.

Решение: предложен следующий алгоритм обработки результатов:

1. Поддерживаются 2 основных типа референсных значений:
   a. интервал вида **x - y**
   b. полуинтервалы <x, >y
2. Для каждого из вышеизложенных типов исходных референсных значений производится приведение к типам a,b.
3. Если в ходе обработки возвращена ошибка – данное значение заменяется средним по обучающей выборке.

- Данные по проведенным тестам ПЦР и ИФА
  - Неунифицируемость формата результатов (**samples_result**):
    - ИФА:

      Выявлены основные 2 формата представления результатов.

      1. Численное представление igg/igm: «nCoV IgM: 0.12\nnCoV IgG: 0»
      2. Категориальное представление igg/igm: «nCoV IgG: Не обнаружено\nnCoV IgM: Не обнаружено» в различных регистровых форматах.

Решение: создание 4х признаков для ИФА: igg_n, igm_n (отражают конкретное значение igg, igm), igg_def, igm_def (отражают бинарный признак обнаружения). На основе igg_n, igm_n также заполняются igg_def, igm_def по следующим правилам:

1. igg_def «Обнаружено», если igg_n > 10.0
2. igm_def «Обнаружено», если igm_n > 2.0

3.1.2.2 Поиск, удаление или исправление артефактов, выбросов и противоречивых данных

- Данные из КТ-центров
  - Несогласованность полей (**temperatura_tela_time** и **event_start_time)**

    Между датой проведения КТ и датой измерения температуры могло пройти много дней. 299’792 наблюдений не имеют данную проблему, однако, среди оставшихся наблюдений, наблюдается следующая картина:
    - 1 день – 2142 наблюдения
    - 2 дня – 266 наблюдений
    - 3 дня – 192 наблюдения
    - 4 дня – 178 наблюдений
    - 5 дней – 95 наблюдений
    - 6 дней – 111 наблюдений
    - 7 дней – 89 наблюдений
    - Более 7 дней – 763 наблюдений (максимальная разница во времени – 176 дня)

Решение: удаление наблюдений с разницей во времени более 7 дней.

  - Несогласованность полей (**KT_stepen_tjazhesti** и **resultat_KT)**

Согласно описанию исходного набора данных, корректно сопоставление между категорией КТ и степенью тяжести КТ: КТ-0 – нулевая, КТ-1 – легкая, КТ-2 – средне-тяжелая, КТ-3 – тяжелая, КТ-4 – критическая.

a. Однако, в наборе данных такому сопоставлению отвечают лишь 167’459 наблюдений. Полная картина представлена в Таблице 1 (**KT_stepen_tjazhesti** и **resultat_KT**):

Таблица 1 — Соответствие категории КТ и степени тяжести КТ

	Нулевая	**Легкая**	**Средне-тяжелая**	**Тяжелая**	**Критическая**
0	536	1423	409	78	0
1	510	157298	2578	126	7
2	48	3208	56516	463	8
3	12	402	1285	15814	9
4	3	30	37	169	1196

Также, в 61463 наблюдениях значение **KT_stepen_tjazhesti** не указано.

Решение: будем считать, что пациент имеет степень тяжести КТ **N**, если его поле **resultat_KT** имеет значение КТ-**N**, а **KT_stepen_tjazhesti** имеет степень тяжести **не ниже** КТ-**N или не указана.**

3.1.3 Формирование проверочных выборок

На вход подаются данные медицинских анализов пациентов из набора ДИТ от 15.02.21 обо всех клинических и амбулаторных COVID-19 пациентах, отфильтрованные описанными в пункте 3.1.2 алгоритмами.

Необходимо сформировать на их основе тренировочную, валидационную и тестовую выборки, предварительно проведя для этого дополнительную фильтрацию, предобработку признаков, балансировку классов, заполнение пропусков и постобработку данных.

С учетом проведенных фильтраций, описанных в пункте 3.1.2, сформированный набор данных имеет следующие характеристики:

1. Объем данных: 403 Мб
2. Количество наблюдений: 299’792
3. Распределение поля resultat_KT (основное поле при формировании целевой переменной) имеет следующий вид (см. Рисунок 1):

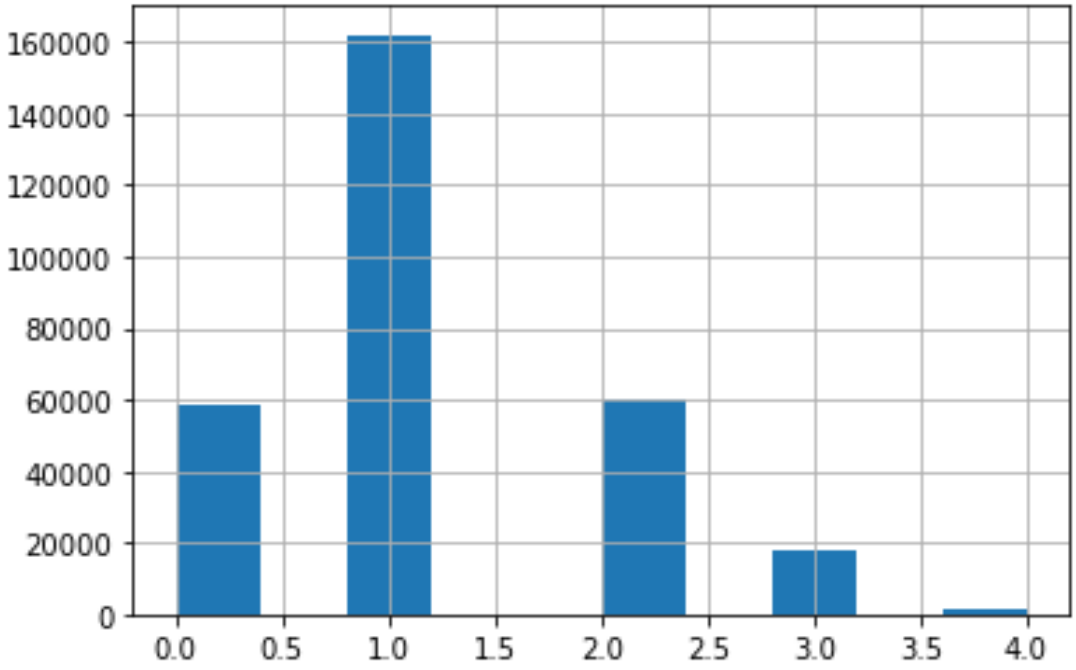

Рисунок 1 — Распределение поля resultat_КТ

Общая картина частот встречаемости КТ выглядит следующим образом:

КТ 1 162071 наблюдение

КТ 2 60485 наблюдений

КТ 0 58226 наблюдений

КТ 3 17568 наблюдений

КТ 4 1442 наблюдения

4. На рисунках 2-7 представлена демографическая статистика набора данных по признакам: сахарный диабет, ожирение, ИБС, артериальная гипертензия, пол, возраст.

Рисунок 2 — Гистограмма сахарного диабета

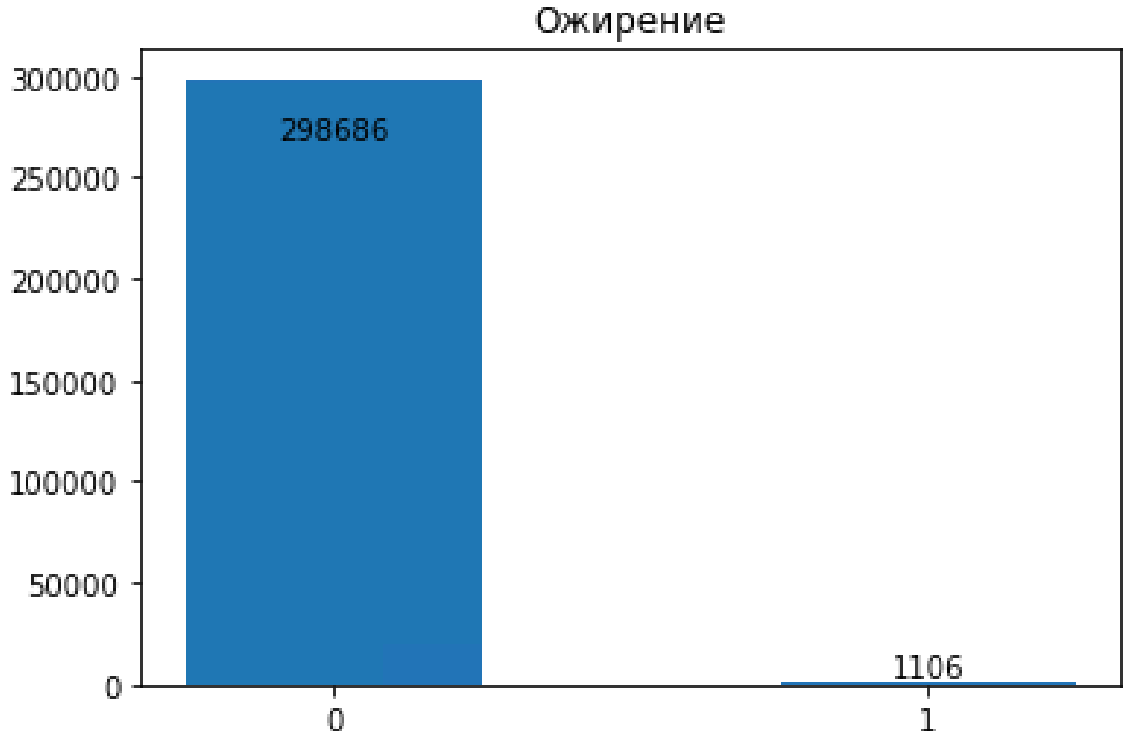

Рисунок 3 — Гистограмма ожирения

Рисунок 4 — Гистограма ИБС

Рисунок 5 — Гистограмма артериальной гипертензии

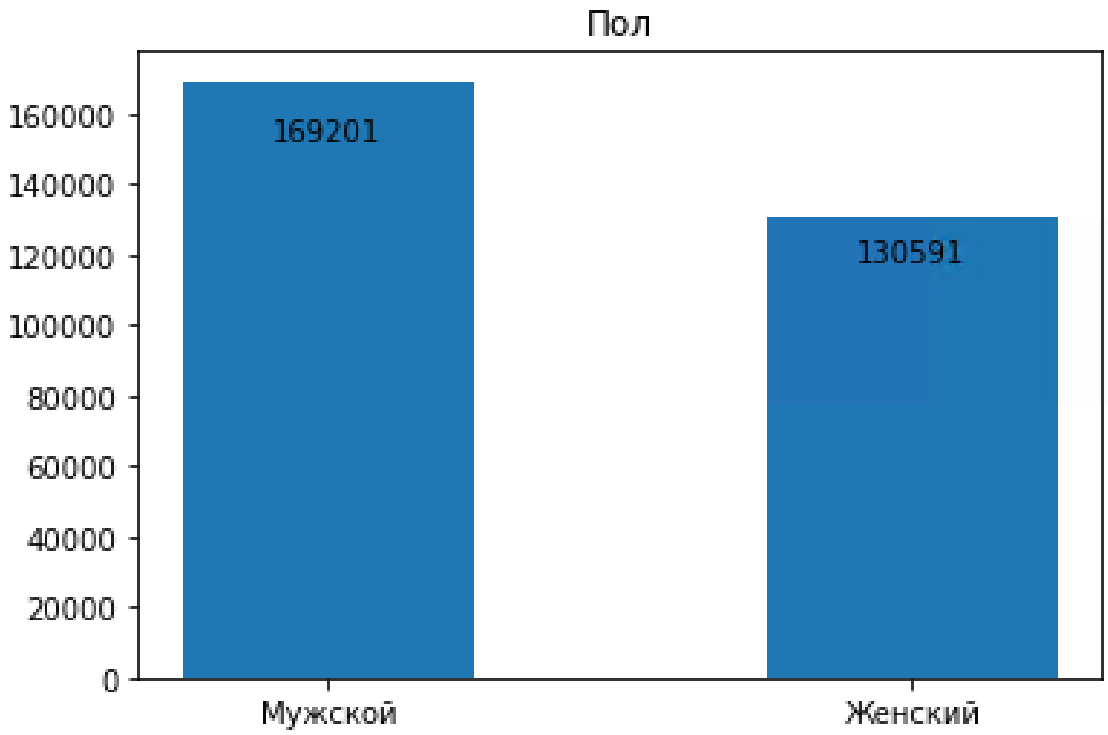

Рисунок 6 — Гистограмма Пола

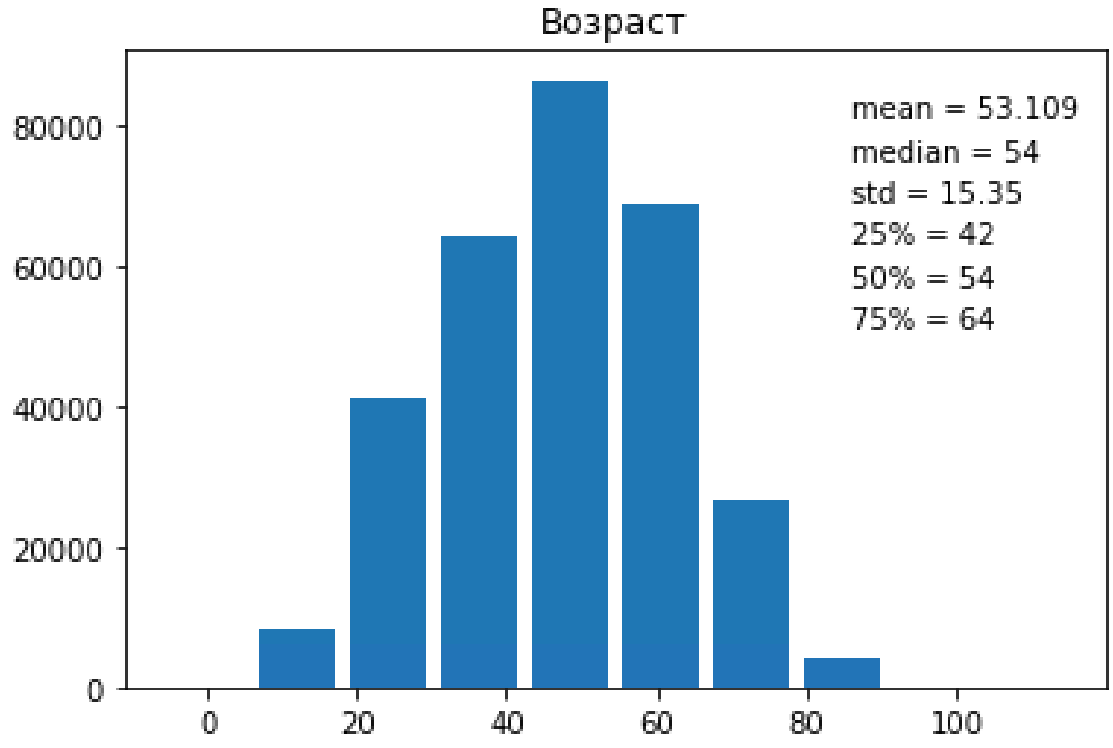

Рисунок 7 — Описательная статистика по возрасту

5. Общий анализ крови (не ограничивая общности, под наличием ОАК будем понимать наличие значения признака «Гематокрит») заполнен на 176’354 наблюдениях.

6. Биохимический анализ крови (будем понимать наличие значения признака «Определение белков острой фазы С-реактивный белок») заполнен на 176’866 наблюдениях.
7. ОАК и биохимический анализ одновременно заполнен на 151’532 наблюдениях.

На основе полученного набора данных формируются следующие выборки:

1. Выборка для обучения классификаторов КТ 01-234 (1, если КТ 234, иначе 0)
2. Выборка для обучения классификаторов КТ 012-34 (1, если КТ 34, иначе 0)

Алгоритм формирования выборок представлен в виде утилиты, инициализируемой константными значениями, и содержит следующие этапы:

1. Фильтрация:

   В ходе данного этапа оставляются только наблюдения с известной целевой переменной, а также, опционально, события с заполненным ОАК и биохимическим анализом крови (как говорилось раннее, под заполненностью понимается наличие признаков «Гематокрит» и «Определение белков острой фазы С-реактивный белок»).
2. Предобработка признаков:

   В ходе данного этапа строковые и категориальные признаки преобразуются к числовому формату согласно внутреннему словарю сопоставлений.

   Также, на данном этапе бинаризуется целевая переменная.

   Для классификатора КТ 01-234 наблюдения с КТ 234 имеют целевое значение 1 (с условием согласованности **resultat_KT** и **KT_stepen_tjazhesti** (будем считать, что пациент имеет степень тяжести из класса КТ **N**, если его поля **resultat_KT** и **KT_stepen_tjazhesti** одновременно принадлежат классу КТ **N))**.

   Целевое значение 0 имеют наблюдения из класса КТ 01 (также, при выполнении условия корректности). Аналогично производится разметка выборки для классификатора КТ 012-34.
3. Балансировка классов:

   На основе целевой переменной унифицируются размеры классов значений. Пусть К – количество наблюдений в меньшем классе. Тогда, для всех классов значений, в которых наблюдений больше чем К, случайным образом выбирается К наблюдений. Полученные множества наблюдений объединяются.
4. Заполнение пропусков:

При указании опции «Заполнить пропуски», в каждом признаке пропуски заменяются на медиану, если признак категориальный, и на среднее значение, если признак непрерывный. Также, в случае заполнения пропусков, для каждого признака N формируется признак none_N, отображающий наличие в оригинальном признаке пропусков (1, если признак имел пропуск, который был заполнен, 0 иначе).

5. Постобработка:

   На данном этапе возможно встраивание дополнительных функций для обогащения выборок (например, формирование признака баллов по шкале NEWS).

6. Разбиение выборок:

   На основе построенных выборок порождаются тестовые, валидационные и обучающие выборки.

   В случае рандомизированного разбиения, пропорциональность наблюдений в выборках: 1/1/8 (то есть в обучающей выборке находится 8/10 наблюдений оригинальной выборки, а в тестовой и валидационной выборках оставшиеся наблюдения распределены в равных долях).

   В случае разбиения по времени, в тестовую выборку определяется первая 1/10 наблюдений за весь период с максимальным event_start_time, а в валидационную выборку последующие 1/10 наблюдений. Остальные наблюдения относятся к обучающей выборке.

Таким образом, построены обучающие, тестовые и валидационные выборки со следующими характеристиками:

- Задача классификации КТ 01-234:

  Общее количество наблюдений: 67’648

  Пропорциональность разбиения: 6’764/6’764/60’884

- Задача классификации КТ 012-34:

  Общее количество наблюдений: 12’576

  Пропорциональность разбиения: 1’257/1’257/10’062

3.1.4 Оценка и доработка моделей-кандидатов на включение в программный сервис

На вход подаются сформированные в разделе 3.1.3 выборки данных медицинских анализов пациентов из набора ДИТ от 15.02.21 обо всех клинических и амбулаторных COVID-19 пациентах.

Необходимо предложить возможные варианты доработки моделей машинного обучения, построенных на предыдущих этапах исследований, а также методики оценки качества работы предложенных моделей.

На основе сформированных обучающих, валидационных и тестовых выборок производилось обучение прогнозных моделей, а также проводилась оценка качества. Для оценки качества прогнозирования будем использовать метрику *ROC-AUC* [10].

Площадь под ROC-кривой (ROC-AUC)

ROC-кривая – график, отображающий соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак, при варьировании порога решающего правила (ошибок I рода).

Площадь под ROC-кривой AUC (англ. Area Under Curve) принимает значение от 0 до 1 и интерпретируема как вероятность того, что классификатор присвоит больший вес случайно выбранному положительному наблюдению, чем случайно выбранному отрицательному наблюдению.

Модель, максимизирующая метрику ROC-AUC, признается лучшей и встраивается в следующую версию Калькулятора КТ. Описание исследуемых моделей и экспериментальные результаты подробно описаны в главе 0 данного отчета.

3.1.5 Выводы

В рамках данного раздела было проведено формирование проверочных выборок для их дальнейшего использования при доработке функционала КТ-калькулятора. Были описаны предоставленные наборы «сырых» данных, а также предложены и реализованы алгоритмы по их предобработке: очистке данных и приведению значений показателей к общим шкалам и словарям, а также по поиску, удалению и исправлению артефактов, выбросов и противоречивых данных. Была разработана методика для формирования набора данных для прогнозирования КТ на основе физикальных и клинических признаков. В частности, были предложены собственные алгоритмы обработки и унификации единиц измерения (в случае, если одна и та же величина в разных наборах данных представлена в разных единицах измерения) и учета референсных значений. Проведен анализ актуальности результатов анализов, в результате которого было принято решение оставлять для дальнейшего анализа только те показания, которые были проведены не ранее недели до даты проведения КТ-исследования. Были предложены алгоритмы фильтрации, предобработки признаков и

балансировки классов результирующей выборки, а также заполнения пропусков, постобработки данных и разбиения выборок на обучающую, валидационную и тестовую части.

Для оценки моделей-кандидатов на включение в сервис «Калькулятор КТ» было предложено использовать площадь под ROC-кривой (ROC AUC). Подробное описание предложенных в рамках данного этапа моделей, а также их экспериментальное сравнение, представлено в следующей главе данного отчета.

3.2 Доработка моделей машинного обучения и выбор лучшей модели для программного сервиса

В качестве моделей прогнозирования степени тяжести использовались модели машинного обучения с учителем [11, 12, 13], в частности, одноклассовые и многоклассовые классификаторы: метод случайного леса [14, 15, 16, 17, 18], нейронные сети [19, 20, 21, 22, 23, 24] и градиентный бустинг [25, 26, 27].

Метод случайного леса (Random Forest)

Основная идея Случайного леса заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт невысокое качество классификации, но за счёт их большого количества результат получается более точным.

Пусть обучающая выборка состоит из N примеров, размерность пространства признаков равна M, и задан параметр m – количество признаков для обучения. Наиболее распространённый способ построения ансамбля деревьев – бэггинг, заключается в следующем:

- Генерируется случайная подвыборка с повторениями размером N из примеров обучающей выборки.
- Построим решающее дерево, классифицирующее образцы сгенерированной подвыборки, причём при добавлении очередного узла дерева будем выбирать набор признаков, на основе которых производится разбиение.
- Дерево строится до полного исчерпания подвыборки.

Итоговая классификация объектов проводится путём голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Нейронные сети (Neural Networks)

Функционирование нейронной сети имитирует функционирование человеческой нейронной системы мозга. Доказано, что с помощью нейронных сетей можно сколь угодно точно аппроксимировать любую непрерывную функцию и имитировать любой непрерывный автомат.

Поскольку модель нейрона реализует функцию от его входов, нейроны можно объединять в соответствии с правилами суперпозиции функций, получая более сложные модели, называемые персептронами или искусственными нейронными сетями прямого распространения.

Для решения задачи прогнозирования был выбран многослойный персептрон, представляющий собой обобщение однослойного персептрона – однослойной нейронной сети, все нейроны которой имеют жесткую пороговую функцию активации.

Многослойный персептрон имеет несколько отличительных признаков: каждый нейрон имеет нелинейную функцию активации, сеть содержит один или несколько слоев скрытых нейронов. Также для многослойного персептрона выделяют два типа сигналов:

1. Функциональный сигнал – это входной сигнал сети, передаваемый по всей сети в прямом направлении. В каждом нейроне, через который передается функциональный сигнал, вычисляется функция активации от взвешенной суммы его входов с поправкой в виде порогового элемента – единичного сигнала с весовым коэффициентом.
2. Сигнал ошибки – это сигнал выхода сети и распространяющийся в обратном направлении от слоя к слою. Сигнал ошибки вычисляется каждым нейроном на основе заданной функции ошибки.

Обучение многослойного персептрона состоит в подборе значений весов слоев сети, чтобы при заданном входном векторе получить на выходе значения сигналов, которые с требуемой точностью будут совпадать с ожидаемыми значениями.

Для обучения многослойного персептрона используется метод обратного распространения ошибки (от англ. Back propagation) – алгоритм обучения, основанный на вычислении градиента функции ошибок. В процессе обучения веса нейронов каждого слоя нейросети корректируются с учетом сигналов, поступивших с предыдущего слоя, и невязки (отклонения) каждого слоя, которая вычисляется рекурсивно в обратном направлении от последнего слоя к первому.

При одноклассовой классификации в качестве функции ошибок использовалась бинарная кросс-энтропия, а в качестве функции активации использовалась логистическая функция. Отметим, что построенная нейросеть является регрессионной, поскольку на ее выходе находится вероятность, а не бинарный ответ.

Градиентный бустинг (Gradient Boosting Machines)

Бустинг – итерационный алгоритм, реализующий «сильный» классификатор, который позволяет добиться произвольно малой ошибки обучения (на обучающей выборке) на основе композиции «слабых» классификаторов.

Основной идеей бустинга является использование весовой версии одних и тех же обучающих данных вместо случайного выбора их подмножества.

«Слабые» классификаторы образуются последовательно, различаясь только весами обучающих данных, которые зависят от точности предыдущих классификаторов. Большие веса назначаются «плохим» примерам, что позволяет на каждой итерации сосредоточиться на примерах, неправильно классифицированных.

Представим функцию f классификации в виде композиции T функций, так что каждая из последующих функций минимизирует остатки от предыдущей. Используя метод градиентного спуска, будем модифицировать модель (так как модификация f на основе отклонений аналогична модификации f на основе отрицательного градиента).

Тогда алгоритм градиентного бустинга может быть описан следующими этапами:

1. Построение исходной модели f(x)
2. В цикле от t = 1 … T (по «глубине» композиции):
    a. Вычисление антиградиента
    b. Построение регрессии h_t по антиградиенту
    c. Модификация модели $f(x) \leftarrow f(x) + \rho_t * h_t(x)$

Преимущество формулировки алгоритма, используя градиент, в том, что можно рассматривать другие функции потерь.

Градиентный бустинг является наиболее общим из всех бустингов, может использовать произвольную функцию потерь и подходит для различных задач (классификация, регрессия, ранжирование).

3.2.1 Раздельное прогнозирование тяжести только по анализу крови с помощью нейросети и только по результатам осмотра с помощью ансамблей деревьев решений (алгоритм машинного обучения «случайный лес») для последующего объединения откликов моделей через ансамбль

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести раздельное прогнозирование степени тяжести КТ только по анализу крови с помощью нейросети и только по результатам осмотра с помощью ансамблей деревьев решений для последующего объединения откликов моделей через ансамбль.

Из признакового пространства выделяются 2 группы признаков. Первая группа – признаки лабораторных анализов. В данную группу входят все показатели из общего анализа крови и биохимических анализов. Остальные признаки отнесем ко второй группе.

На основе групп признаков набор данных разделяется на два поднабора. Набор данных, основанный на признаках лабораторных анализов, используется для обучения нейросетевой модели. Отклик модели добавляется во второй набор данных как отдельный признак. На основе дополненного второго набора производится обучение модели Случайный лес. Вероятностный прогноз Случайного леса используется при оценке качества модели. Также, модель может быть дополнена калибровкой Платта.

На рисунках 8-11 изображены ROC-AUC кривые классификаторов 01-234 и 012-34 на рандомизированной и временной выборке.

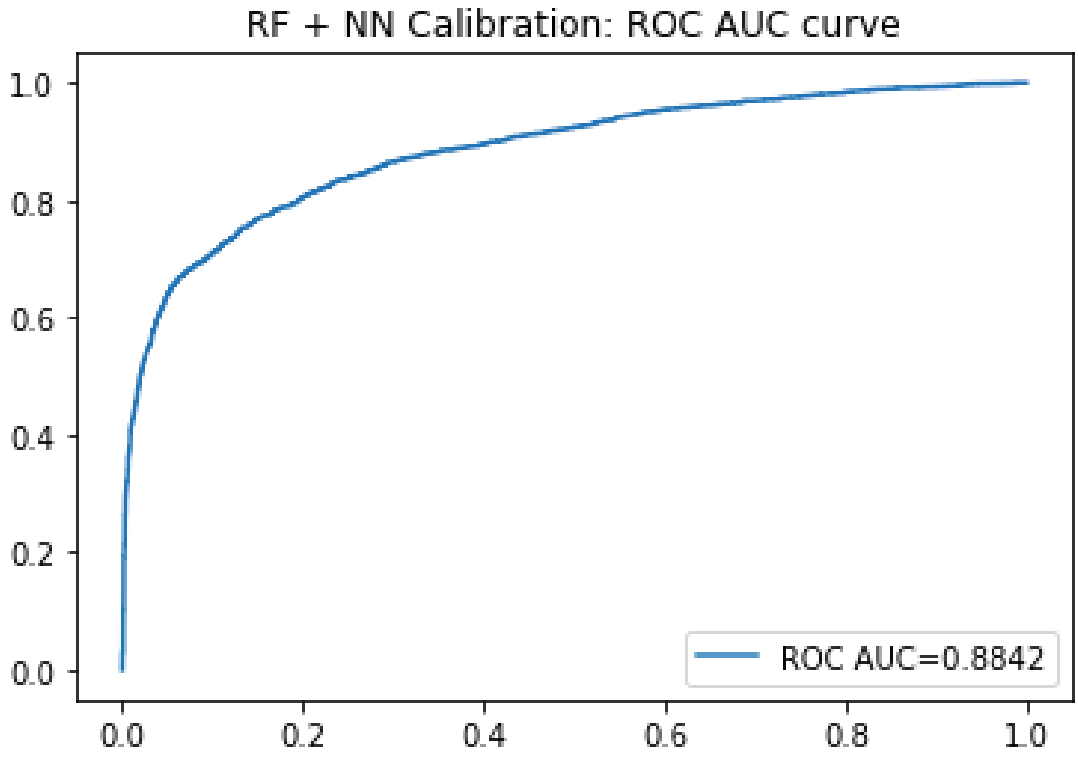

Рисунок 8 — RF+NN 01-234 Рандомизированная выборка

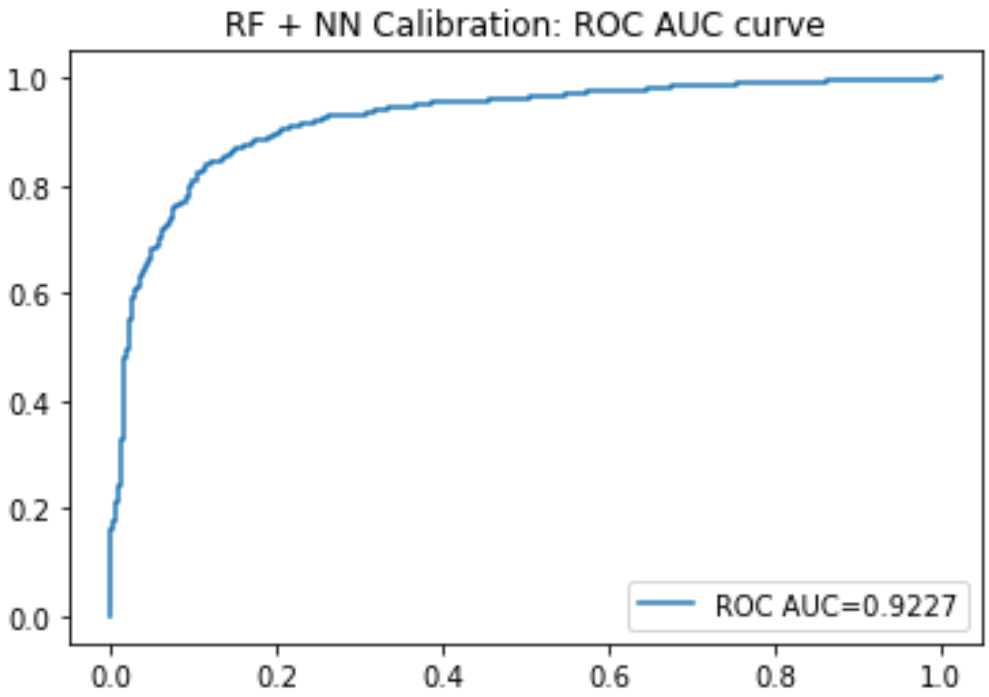

Рисунок 9 — RF+NN 012-34 Рандомизированная выборка

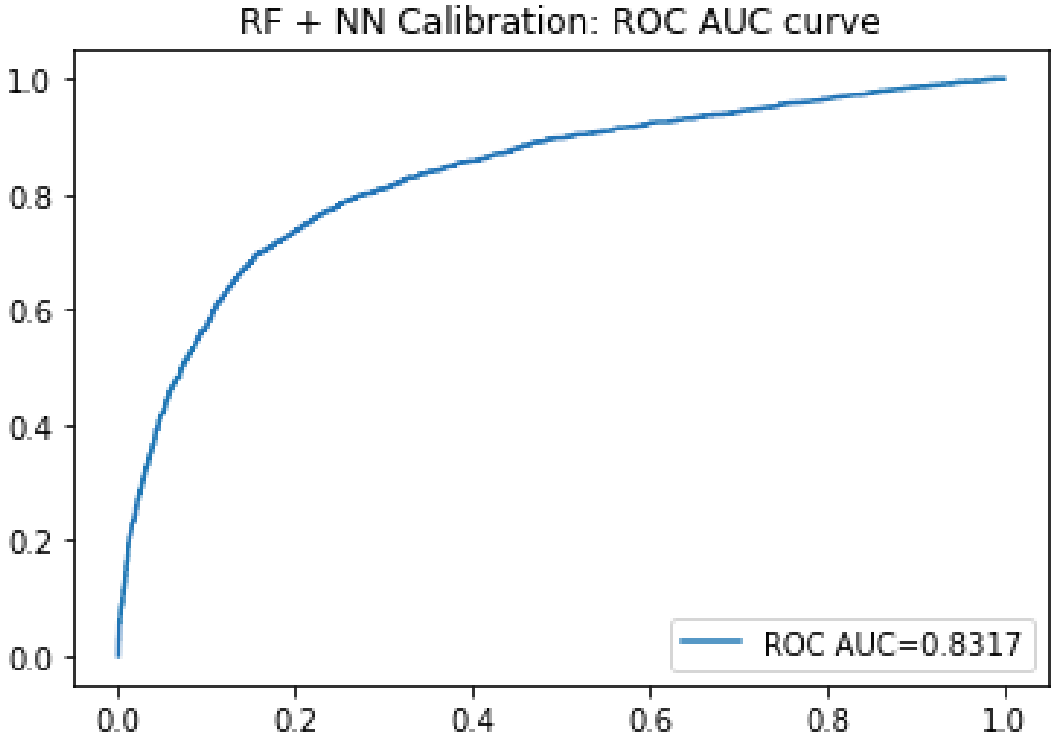

Рисунок 10 — RF+NN 01-234, выборка по времени

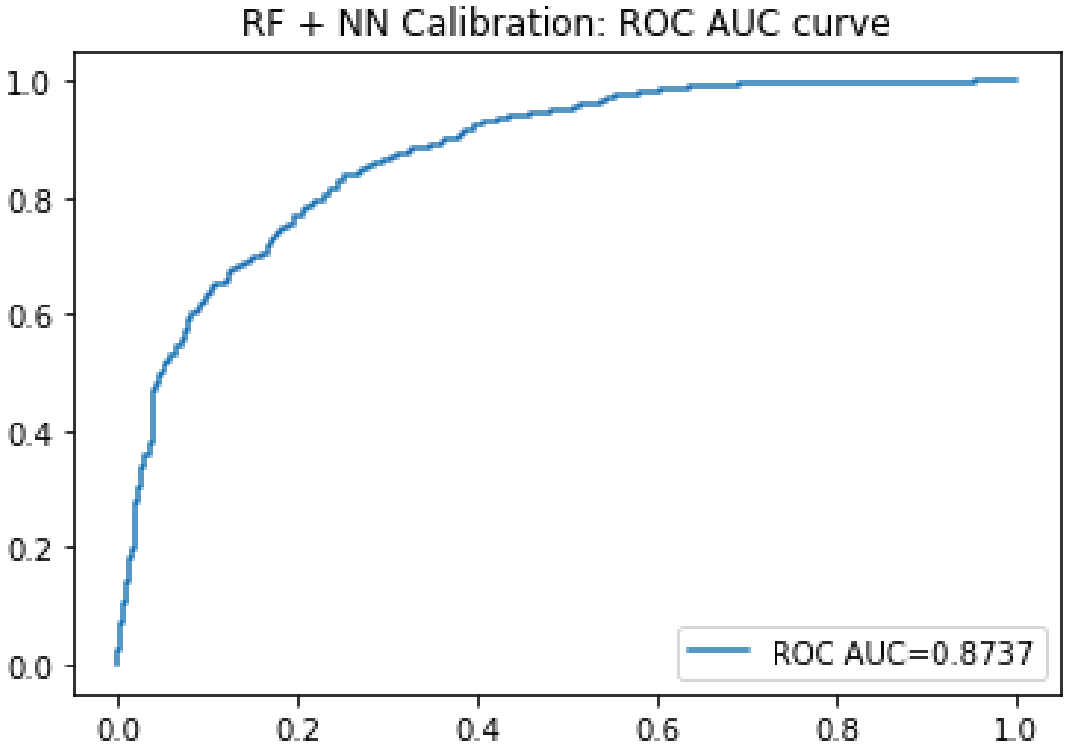

Рисунок 11 — RF+NN 012-34, выборка по времени

3.2.2 Использование бустинг ансамбля деревьев решений lgbm (алгорим машинного обучения «градиентный бустинг») для прогнозирования степени тяжести КТ

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести прогнозирование степени тяжести КТ с использованием бустинг ансамбля деревьев решений LGBM.

LGBM (Light Gradient Boosting Machine) [28, 29] – это одна из наиболее эффективных реализаций процедуры градиентного бустинга. Поскольку данный метод относится к классу ансамблевых алгоритмов машинного обучения, где ансамбли строятся на основе деревьев решений, LGBM позволяет оценивать важность признаков из обученной модели. Как правило, важность обеспечивает оценку, которая указывает, насколько полезным был каждый признак при построении деревьев решений в модели. Чем больше атрибут используется для принятия ключевых решений, тем выше его относительная важность.

Важность рассчитывается для отдельного дерева решений, затем значения характеристик усредняются по всем деревьям решений в модели.

Особенность LGBM заключается в использовании градиентной односторонней выборки (GOSS) и объединения взаимоисключающих признаков (EFB). Обучение происходит только на тех данных, которые приводят к большему градиенту, что способствует ускорению работы алгоритма и уменьшению его вычислительной сложности. EFB позволяет автоматически объединять разреженные (в основном, нулевые) взаимоисключающие признаки, такие как категориальные переменные входных данных.

На рисунках 12-19 изображены ROC-AUC кривые классификаторов 01-234 и 012-34 на рандомизированной и временной выборке.

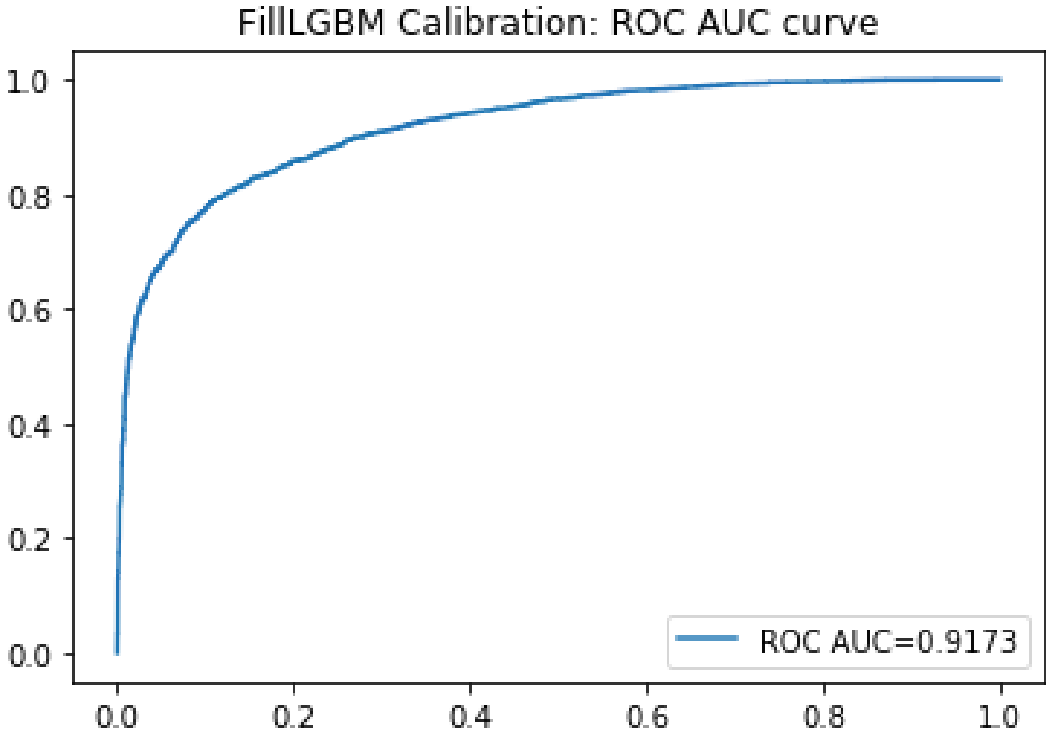

Рисунок 12 — LGBM 01-234 с заполнением пропусков, рандомизированная выборка

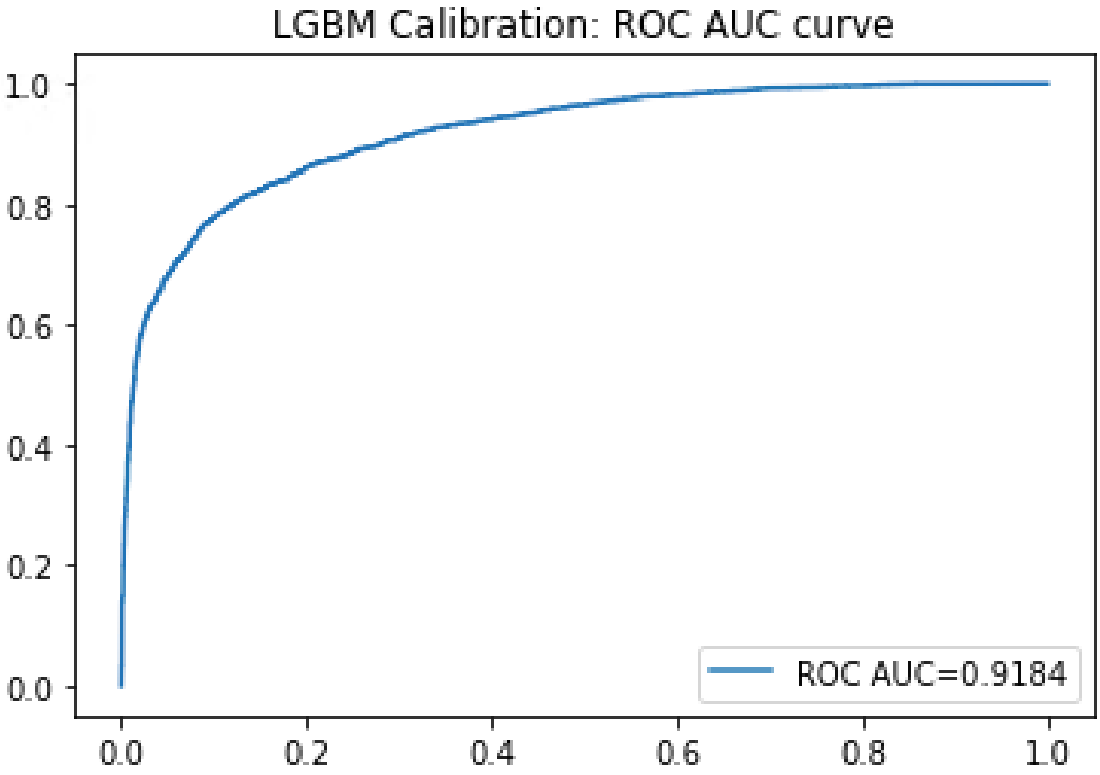

Рисунок 13 — LGBM 01-234 без заполнения пропусков, рандомизированная выборка

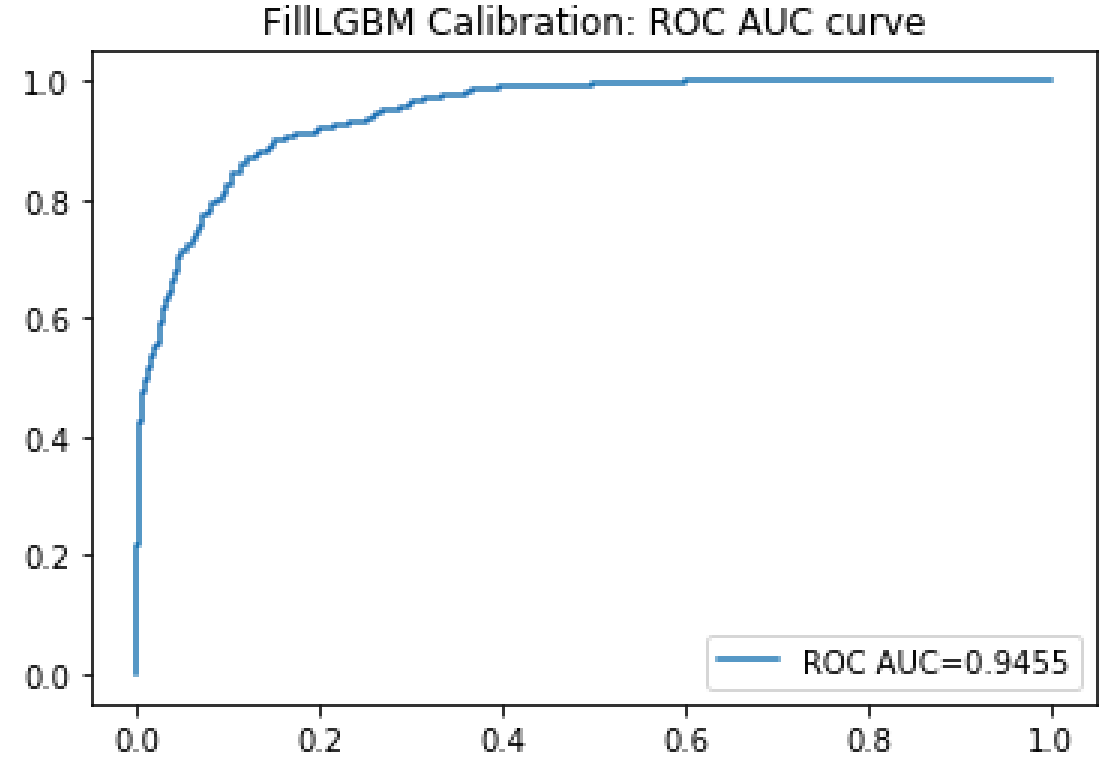

Рисунок 14 — LGBM 012-34 с заполнением пропусков, рандомизированная выборка

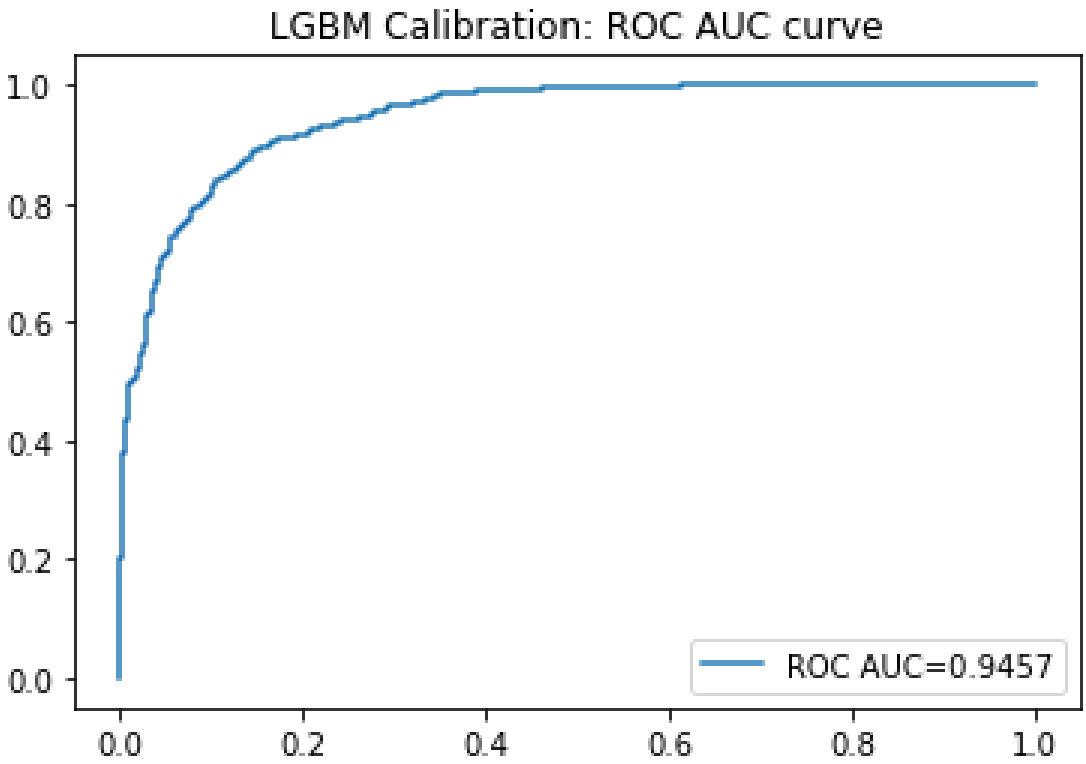

Рисунок 15 — LGBM 012-34 без заполнения пропусков, рандомизированная выборка

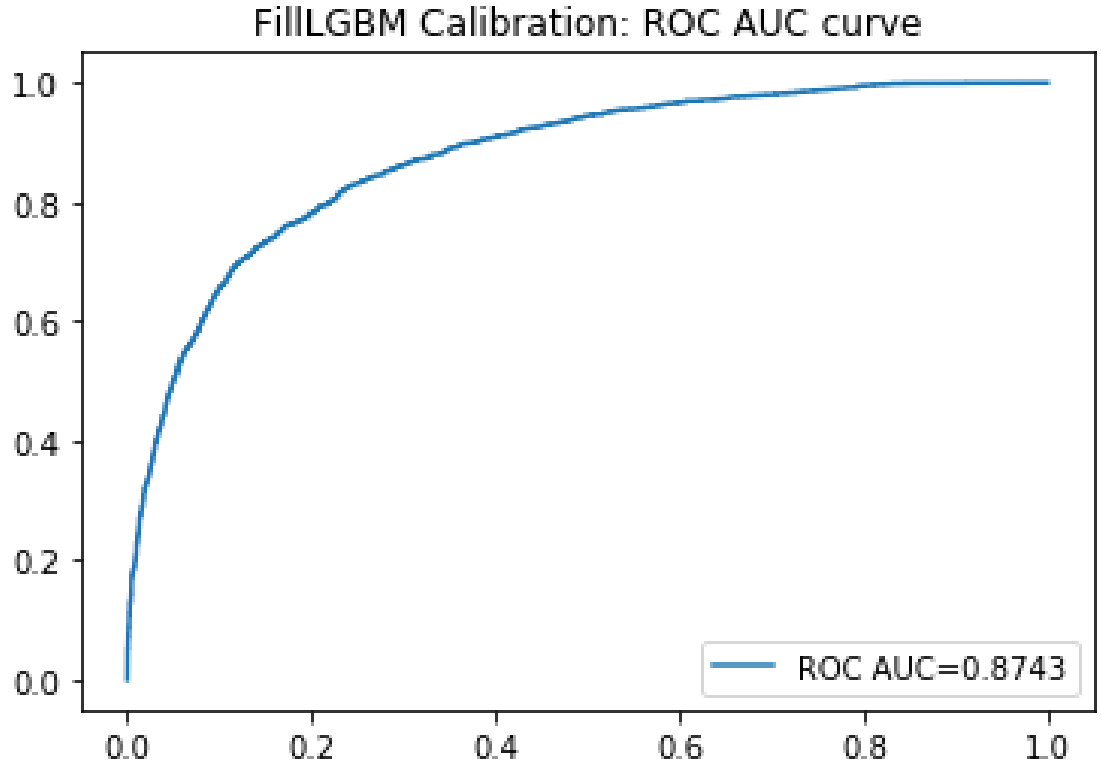

Рисунок 16 — LGBM 01-234 с заполнением пропусков, выборка по времени

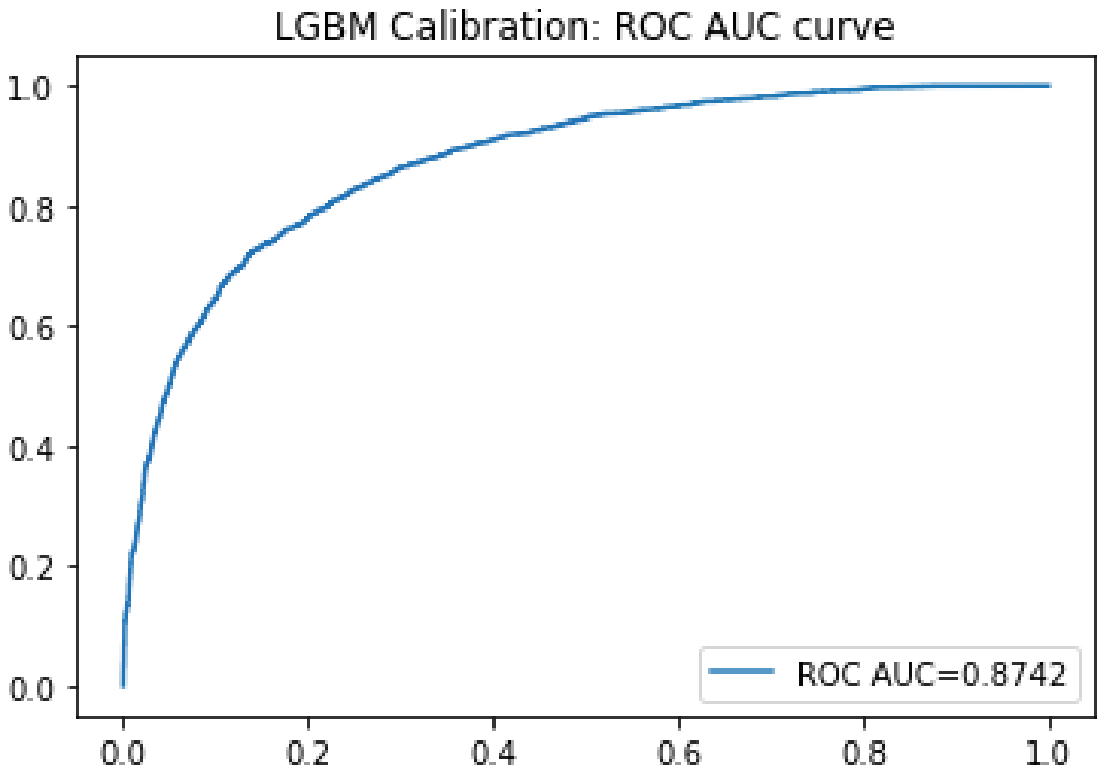

Рисунок 17 — LGBM 01-234 без заполнения пропусков, выборка по времени

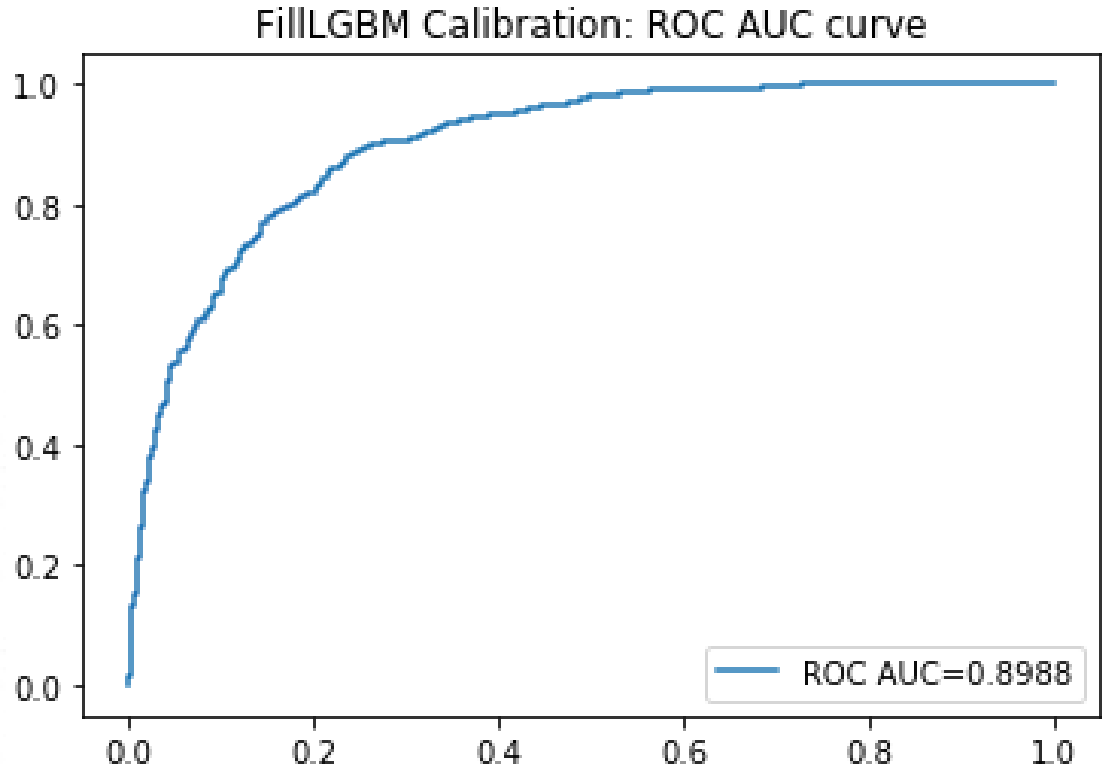

Рисунок 18 — LGBM 012-34 с заполнением пропусков, выборка по времени

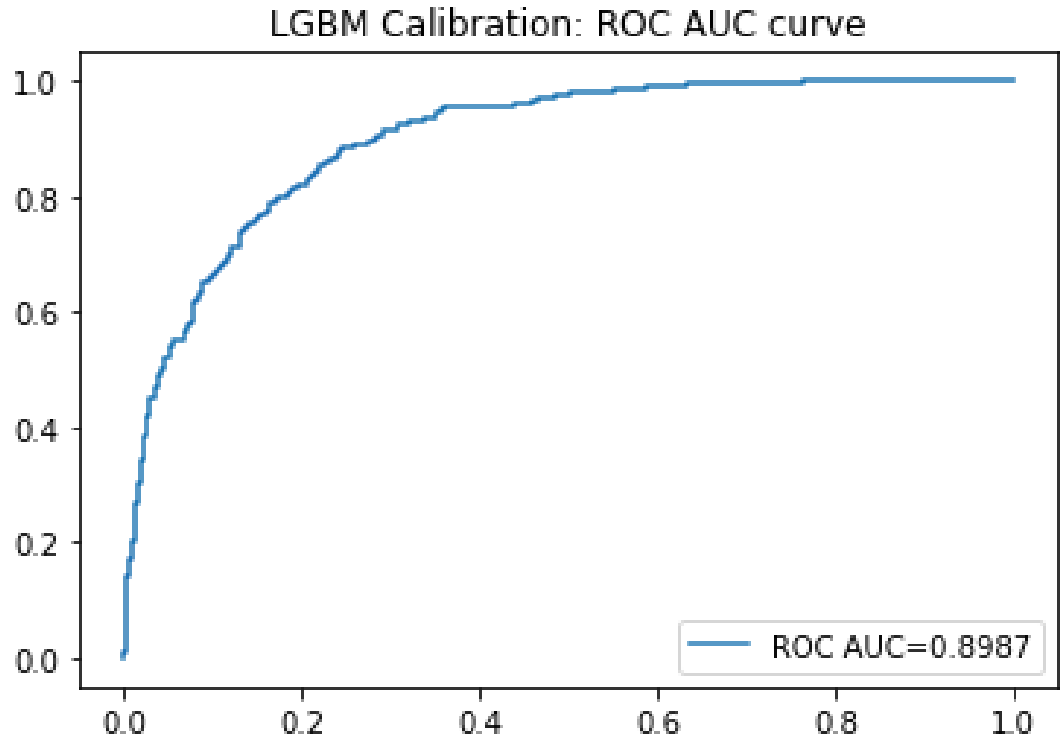

Рисунок 19 — LGBM 012-34 без заполнения пропусков, выборка по времени

3.2.3 Использование ансамбля регуляризованных нейросетей для прогнозирования степени тяжести КТ

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести прогнозирование степени тяжести КТ с использованием ансамбля регуляризованных нейросетей.

В чистом виде метод обратного распространения ошибки работает плохо [30, 31, 32]. Возникают проблемы медленной сходимости или расходимости, застревания в локальных минимумах функционала. Для стабилизации процесса обучения была добавлена инициализация и регуляризация слоев нейросети. Так как в качестве функции активации нейросети используется сигмоида, отклик нейросети может быть интерпретирован как вероятность.

Однако, при различных random_state, различаются как результаты сходимости нейросети, так и порог бинаризации отклика. Для стабилизации работы частных нейросетей, был предложен подход формирования ансамбля нейросетей. Для входного наблюдения

рассчитывается множество откликов по всем обученным нейросетям, отклики объединяются суммированием и нормируются. Полученный отклик принимается за отклик ансамбля.

На рисунках 20-23 изображены ROC-AUC кривые классификаторов 01-234 и 012-34 на рандомизированной и временной выборке.

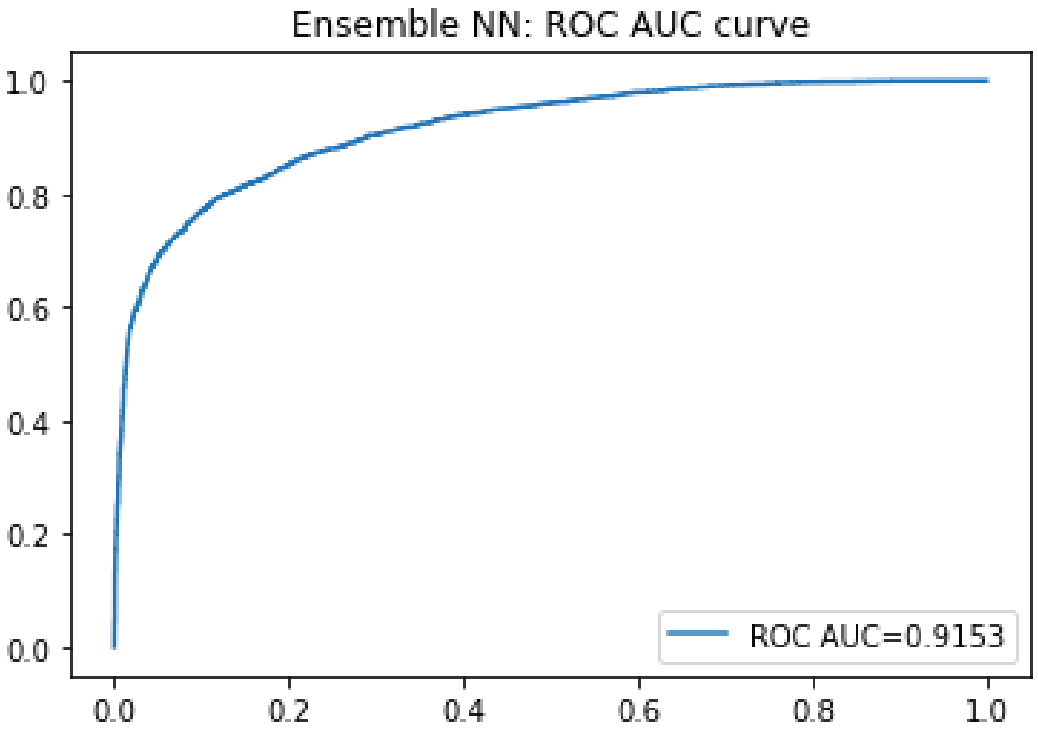

Рисунок 20 — Ансамбль NN 01-234, рандомизированная выборка

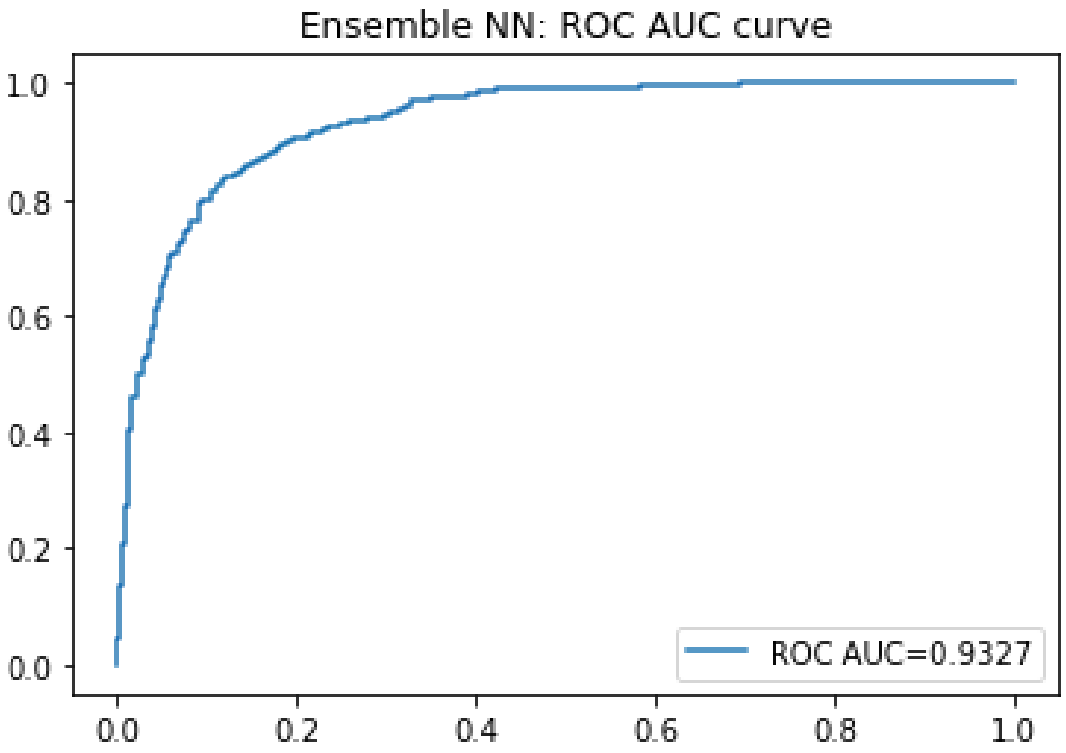

Рисунок 21 — Ансамбль NN 012-34, рандомизированная выборка

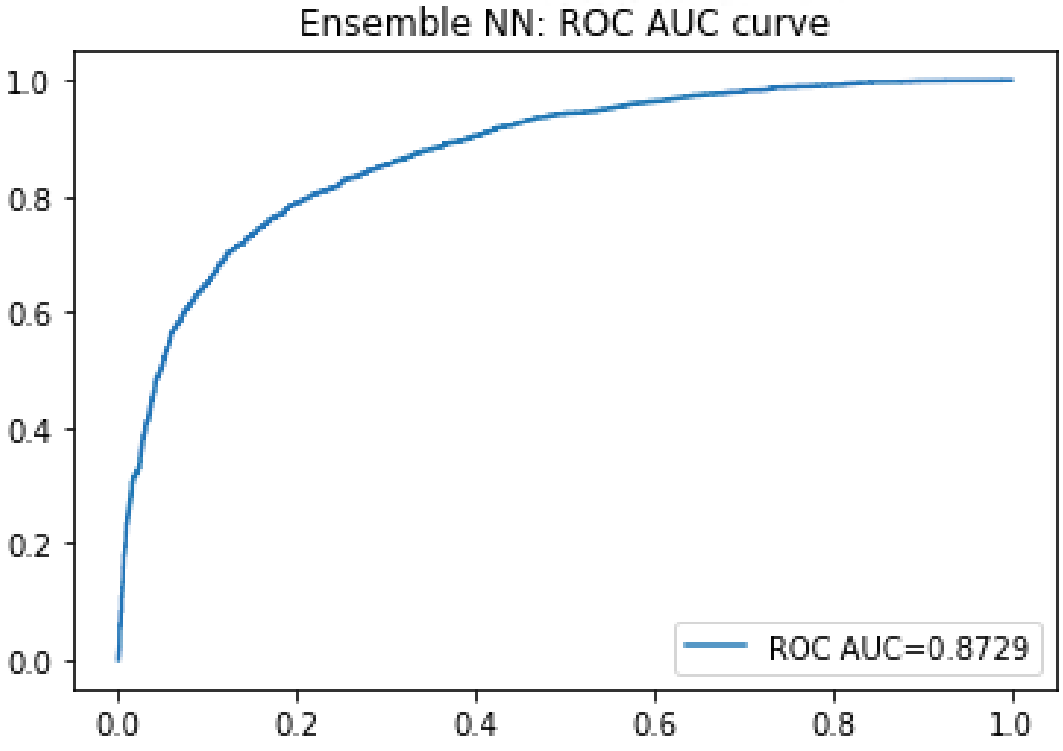

Рисунок 22 — Ансамбль NN 01-234, выборка по времени

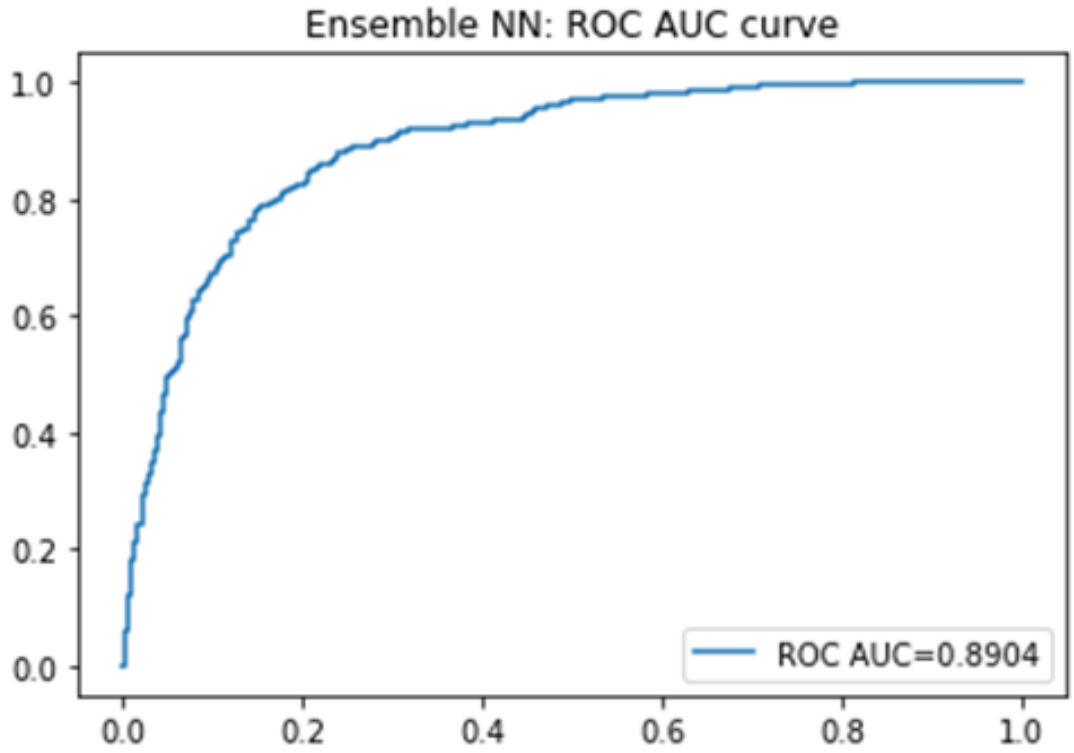

Рисунок 23 — Ансамбль NN 012-34, выборка по времени

3.2.4 Дополнительная очистка и нормализация тренировочной выборки с удалением противоречий

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести дополнительную очистку и нормализацию тренировочной выборки с удалением противоречий.

В разделе 3.1.2 описывалась подготовка данных, проводимая в ходе формирования итогового набора данных. Однако, по словам экспертов, в сформированном наборе существуют следующие противоречия:

1. Неприводимые единицы измерения

   Противоречие было замечено в следующих признаках:

   a. Абсолютное количество эозинофилов
   b. Абсолютное количество базофилов
   c. Абсолютное количество моноцитов
   d. Абсолютное количество гранулоцитов
   e. Абсолютное количество лимфоцитов
   f. Абсолютное количество нейтрофилов

Противоречие заключается в наличии 2х категорий значений: процентном и количественном содержании, а также невозможности преобразования друг в друга общим подходом преобразования единиц измерения (преобразование сокращений измерений в единообразный формат). Для корректного преобразования требуется умножить значения процентного содержания на количество лейкоцитов (WBC). При их отсутствии у наблюдения, значение данных признаков принимается пустым.

2. Отсутствие единиц измерения
    a. RDW
    b. PDW
    c. D-димер

Данные признаки не содержат единиц измерения, поэтому унифицирование значений не может быть обеспечено общим подходом преобразования единиц измерения. Для устранения данной проблемы было принято решение анализировать референсные значения признаков:

a. RDW [33]:

   Если референсная правая граница меньше 20 или написание содержит запятую, значение представлено в виде **RDW-CV** и измеряется в %, иначе представлено в виде **RDW-SD** и измеряется в фемтолитрах.

   Сопоставление единиц измерения производится по правилу:

   RDW-CV=(RDW-SD/MCV)*100, MCV - средний объем эритроцита

b. PDW [34]:

   Если референсная правая граница меньше 20 или написание содержит запятую, значение представлено в виде **PDW-CV** и измеряется в %, иначе представлено в виде **PDW-SD** и измеряется в фемтолитрах.

   Сопоставление единиц измерения производится по правилу:

   PDW-CV=(PDW-SD/MPV)*100, MPV - средний объем тромбоцитов

c. D-димер [35]:

   Если референсная правая граница меньше 1, то значение измеряется в **мкг/мл,** иначе измеряется в **нг/мл.**

   Сопоставление единиц измерения производится по правилу:

   1 **мкг/мл** = 1000 **нг/мл**

3.2.5 Расширение признакового пространства, в том числе за счет включения информации о датах проведения анализов и осмотра

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести расширение признакового пространства, в том числе за счет включения информации о датах проведения анализов и осмотра.

Как было описано в главе 3.1.2, при формировании набора данных наблюдения с проведенной КТ обогащаются клиническими анализами, тестами ПЦР и ИФА, а также сатурацией, взятыми за период +- неделя с даты проведения КТ.

На практике, данные анализов не всегда являются свежими. Для контроля свежести данных были введены 2 признака: дата ОАК и дата биохимического анализа. На их основе вычисляется количество дней между анализами и датой КТ. Признаки оказались значимыми для моделей прогнозирования КТ, поэтому набор был обогащен датами проведения всех анализов с последующим расчетом количества дней.

На основе данных признаков можно проводить фильтрацию актуальности входных данных, а при количестве дней, меньшем 7, использовать признак в прогнозной модели.

3.2.6 Взаимообратный учет показателя сатурации в оценке степени тяжести с пересчетом по баллам шкалы NEWS

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести взаимообратный учет показателя сатурации в оценке степени тяжести с пересчетом по баллам шкалы NEWS.

При прогнозировании ожидаемой степени тяжести КТ одним из важнейших факторов является показатель сатурации (насыщение крови кислородом в процентном соотношении). Однако, в сформированном наборе данных признак сатурации заполнен лишь на 103’458 наблюдениях (из 299’792).

Для увеличения заполненности были сформированы признаки b_spo2, bw_spo2:

a. b_spo2 – баллы за сатурацию по шкале NEWS [36] (приводятся по формуле (97-saturation)/2
b. bw_spo2 – баллы за сатурацию, основанные на заполненных признаках по модифицированной шкале NEWS [36]. Признак формируется на основе ЧДД, температуры тела и степени тяжести осмотра пациента.

Для данных параметров описывается весовая схема, основанная на баллах за степень тяжести осмотра.

Далее, в зависимости от значений ЧДД и температуры тела, из баллов за степень тяжести вычитается некоторое число баллов.

Итоговое количество баллов является ожидаемым значением баллов за сатурацию. Таким образом, для каждого наблюдения ставится балл за сатурацию при её наличии и ожидаемые баллы при отсутствии.

3.2.7 Алгоритмы для автоматизированного тьюнинга и регуляризации моделей

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо рассмотреть алгоритмы для автоматизированного тюнинга и регуляризации моделей.

Использование нейросетевых моделей неизбежно приводит к необходимости перебора большого количества вариаций гиперпараметров. Для поиска оптимальных гиперпараметров и контроля обучения использовались следующие техники:

1. Grid search (перебор параметров по заданной сетке) [37]

   В качестве параметров отбирались:

   - количество слоев NN: от 1 до 4 с шагом 1
   - количество нейронов в одном слое: от 10 до 400 с шагом 10
   - функции активации: linear, sigmoid, tahn, relu
   - оптимизаторы: adam, nadam, SGD, RMSprop
   - функции потерь: mean_squared_error, mean_absolute_error, categorical_hinge, binary_crossentropy, kullback_leibler_divergence

2. Система callbacks при обучении

   - EarlyStopping – обработчик прекращения обучения, когда метрика перестала улучшаться (в данной задаче использовалось значение функции потерь на валидационной выборке). Максимальное количество эпох без возможного улучшения параметра val_loss – 10 эпох.
   - ReduceLROnPlateau – обработчик автоматического уменьшения скорости обучения (lr – learning rate), когда метрика перестала улучшаться (аналогично используется val_loss).

3.2.8 Исследование различных видов калибровок результата и выбора порогов

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести исследование различных видов калибровок результата и выбора порогов.

В данной работе рассматриваются 2 основных подхода калибровки [38, 39, 40, 41] вероятностей прогнозных моделей:

1. Калибровка Платта [40]

   Пусть прогнозная модель выдаёт вероятность p(x), тогда ищется решение вида: $p_{new}(x) = sigmoid(a * p(x) + b)$

   Параметры a, b определяются методом максимального правдоподобия на отложенной выборке. Калибровка Платта наиболее эффективна, когда искажение в предсказанных вероятностях имеет сигмовидную форму.

2. Изотоническая регрессия [41]

   Является более мощным методом калибровки, который может исправить любое монотонное искажение. Анализ кривой обучения показывает, что изотоническая регрессия более склонна к переобучению и, следовательно, работает хуже, чем масштабирование Платта, когда данных недостаточно.

Калибровка моделей [38, 39] нужна для правильного понимания, насколько результатам модели можно доверять. Это важно как при интерпретации моделей, так и для принятия решений о внедрении моделей и анализа их работы.

Также, возникает проблема соотнесения наблюдений к классам на основе спрогнозированной вероятности. В действительности, не все построенные модели имеют порог 0.5 для выходных вероятностей, а также, веса ошибок при «занижении» или «завышении» результата могут отличаться. В данной задаче, вес ошибки «занижения» выше веса ошибки «завышения» и необходимо максимизировать метрику Recall (полнота) или Sensitivity (чувствительность).

Экспертами может быть установлено минимальное допустимое значение recall, на основе которого могут быть найдены пороги [42, 43, 44, 45], максимизирующие precision (точность). Данный подход позволяет использовать множество частных порогов в различных организациях использования Калькулятора КТ, однако не имеет унифицируемого подхода расчета порогов.

3.2.9 Методы непротиворечивого сведения откликов двух независимых бинарных моделей, прогнозирующих отдельно вероятность тяжелого поражения (КТ34) и легкого поражения (КТ01) к многоклассовой вероятностной модели (КТ01, КТ2, КТ34) и к единой шкале оценки тяжести

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо рассмотреть методы непротиворечивого сведения откликов двух независимых бинарных моделей, прогнозирующих отдельно вероятность тяжелого поражения (КТ34) и легкого поражения (КТ01) к многоклассовой вероятностной модели (КТ01, КТ2, КТ34) и к единой шкале оценки тяжести.

На основе построенных моделей прогнозирования вероятности степени тяжести КТ 01 и КТ 34, для входного наблюдения могут быть получены 2 отклика моделей: k234, k34.

На основе данных откликов по формуле полной вероятности могут быть вычислены вероятности каждого класса КТ: 01, 2, 34 следующим образом:

- $P_{КТ\ 01} = \frac{(1-k234)*(1-k34)}{1-k34+k234*k34}$
- $P_{КТ\ 2} = \frac{k234*(1-k34)}{1-k34+k234*k34}$
- $P_{КТ\ 34} = \frac{k234*k34}{1-k34+k234*k34}$

Для расчета ожидаемой степени тяжести КТ (Е) в данной работе используется весовая схема: $E = 1 * P_{КТ\ 01} + 2 * P_{КТ\ 2} + 3 * P_{КТ\ 34}$. Ожидаемую степень можно соотносить к целочисленной степени тяжести на основе неравномерных порогов:

- КТ 01 при $E \leq 1.5$
- КТ 2 при $1.5 \leq E \leq 2.5$
- КТ 34 при $E \geq 2.5$

Однако, подход на основе порогов является корректным не для всех задач. Как говорилось ранее, в данной задаче ошибка «занизить степень тяжести» является критичнее завышения. Также, при максимизации метрики Recall (полнота) или Sensitivity (чувствительность) пороги могут быть сдвинуты вправо (увеличены).

Для оценки уверенности в прогнозе может быть построена единичная шкала степеней тяжести КТ. В таком случае, за интервал от 0.0 до 0.33 отвечает КТ 01, за интервал от 0.34 до 0.66 отвечает КТ 2, за интервал от 0.67 до 1.0 отвечает КТ 34.

Для отображения ожидаемой степени тяжести на шкалу степеней тяжести необходимо сместить ожидаемую степень тяжести на интервал от 0.0 до 1.0 по формуле $\frac{E-1.0}{2.0}$. Однако, в таком случае граничные пороги вероятностей соответствуют 0.25 и 0.75 и остается проблема неравномерности вероятностных интервалов. Для решения проблемы необходимо сместить пороги на 0.33 и 0.66. Такое отображение отражает линейная функция y = 2/3*x+1/6.

Полученное вещественное число может быть отмечено на шкале от 0.0 до 1.0 и визуально отображать, к какой степени тяжести тяготеет наблюдение.

3.2.10 Разработка окружения и скриптов для автоматизированного тестирования сервиса «Калькулятор КТ» через интерфейс REST API

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести разработку окружения и скриптов для автоматизированного тестирования обновленного сервиса «Калькулятор КТ» через интерфейс REST API (с учетом разработанных в рамках данного этапа алгоритмов).

Для предложенных моделей была проведена оценка качества. Таблица 2 содержит информацию о полученных значениях метрики ROC AUC на рандомизированной и временной тестовой выборке для задач классификации КТ 01-234 и КТ 012-34.

Таблица 2 — Таблица значений ROC AUC прогнозных моделей

	КТ 01-234 рандомизированная выборка	**КТ 012-34 рандомизированная выборка**	**КТ 01-234 выборка по времени**	**КТ 012-34 выборка по времени**
NN на лабораторных признаках	0.7954	0.8401	0.7899	0.7984
NN на анализах крови + RF + Калибровка Платта	0.8842	0.9227	0.8317	0.8737
Ансамбль нейросетей	0.9153	0.9327	0.8729	0.8904
LGBM с заполнением пропусков + калибровка Платта	0.9173	0.9455	0.8743	0.8988
LGBM без заполнения пропусков + калибровка Платта	0.9170	0.9453	0.8742	0.8987

Исходя из таблицы результатов, наиболее перспективной для внедрения в Калькулятор КТ является модель LGBM с заполнением пропуском и калибровкой Платта. Данная модель была встроена в тестовую версию Калькулятор КТ. На основе описанных в пунктах 3.1.1.1, 3.1.1.3 наборов данных, может быть проведено тестирование модели на внешних данных. Для проведения тестирования было создано окружение, содержащее как данные для тестирования, так и скрипт автоматического тестирования.

Скрипт написан на языке Python и предоставляет возможность получить прогноз для «сырых» внешних данных. Скрипт состоит из нескольких программных блоков. Блок импорта и предобработки наборов данных обеспечивает корректную загрузку наборов данных в формате xlsx, csv с последующим сопоставлением названий признаков исходного набора данных и признаков сервиса Калькулятор КТ. Далее, в блоке обработки значений набора данных, значения признаков приводятся к шкалам, на которых производилось обучение прогнозной модели. После этого, наблюдения разбиваются на независимые словари значений и отправляются сервису Калькулятор КТ через интерфейс REST API [46, 47].

Архитектурный стиль REST функционирует поверх протокола HTTP. Для каждой операции сопоставляется свой собственный HTTP метод:

- GET – получение данных в удобном для клиента формате;
- POST – создание новых данных;
- PUT – обновление, модификация данных;
- DELETE – удаление данных.

В качестве пакета данных отправляется JSON массив на указанный конкретный URL (адрес сервиса). Со стороны сервиса Калькулятор КТ срабатывает функция-обработчик, а в зависимости от отправленных данных и текущего запроса возвращается прогноз в определенном формате.

После получения прогноза, скриптом проверяется правильность результата, и в случае ошибки, пользователю скрипта выводится информация о настоящей степени тяжести, ошибочном прогнозе и вероятностных характеристиках прогноза. В ходе получения прогнозов формируется выходная таблица по всем наблюдениям. Таблица содержит уникальный идентификатор, исходную разметку и прогноз модели.

На основе разработанного скрипта была проведена апробация прогнозной модели на внешних данных, описанных в пункте 3.1.1.1. На задаче классификации КТ 01-234 получен

ROC AUC, равный 0.8805, а на задаче классификации КТ 012-34 ROC AUC равен 0.9311. ROC AUC кривые классификаторов изображены на Рисунках 24 и 25.

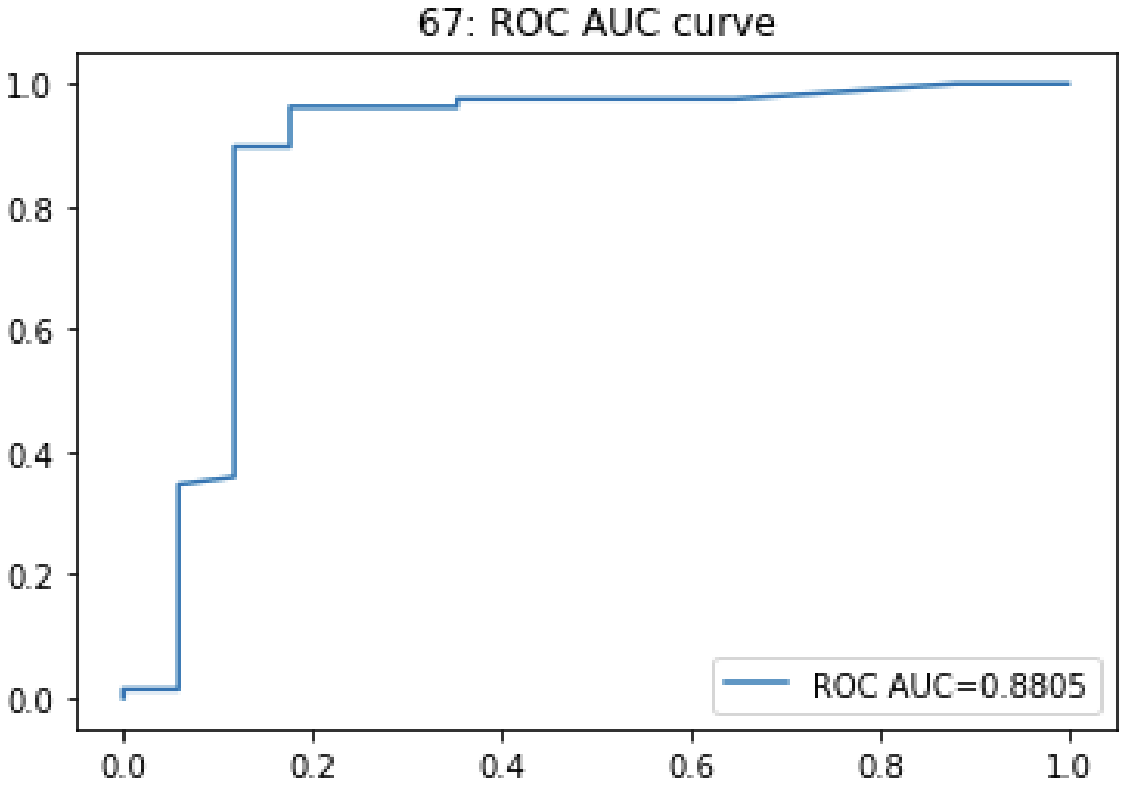

Рисунок 24 — ROC AUC кривая задачи КТ 01-234 на наборе данных ГКБ № 67

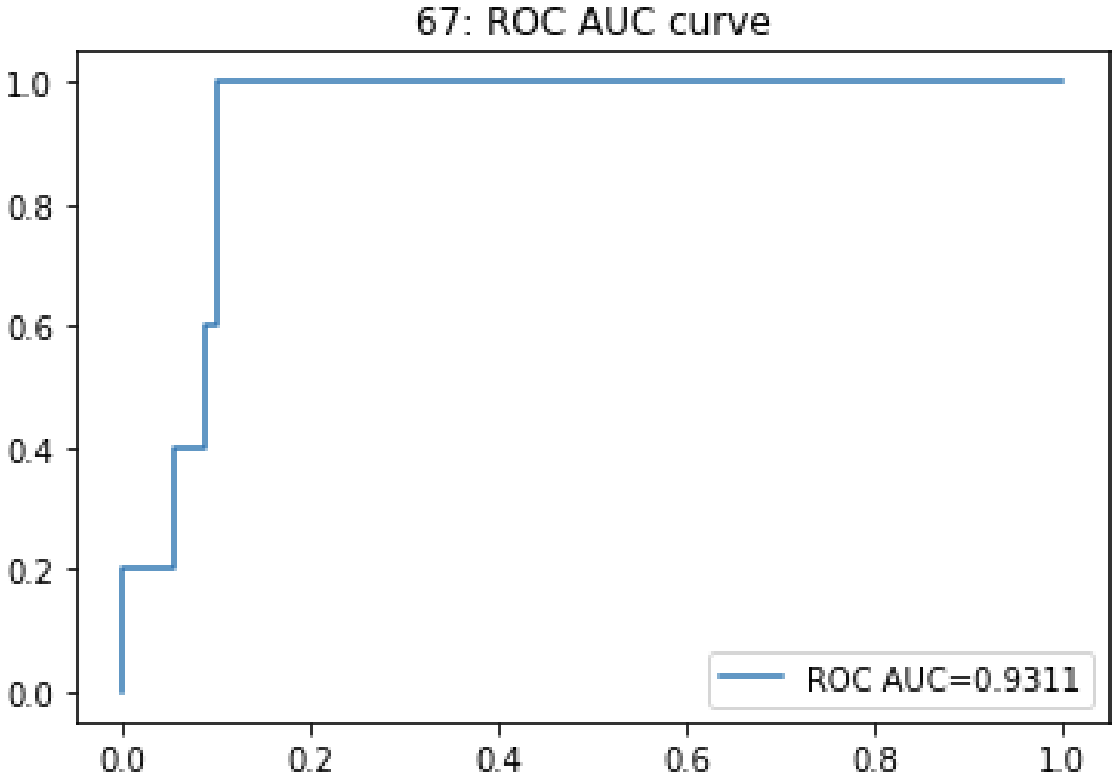

Рисунок 25 — ROC AUC кривая задачи КТ 012-34 на наборе данных ГКБ № 67

На данных, описанных в пункте 3.1.1.3, модель показала низкую точность (сильное «занижение» ожидаемой степени тяжести). Причиной низкой точности является отсутствие наиболее значимых признаков (ЧДД, сатурация, температура тела) в наборе данных. Однако, после ручной проверки исходных данных экспертами, и сама разметка набора данных была признана некорректной.

3.2.11 Разработка средств «объясняющей» визуализации результатов прогноза на основе методов LIME и kernel SHAP, применение и анализ результатов этих методов к ошибкам прогноза

На вход подаются сформированные в разделе 3.1 выборки данных медицинских анализов больных COVID-19 пациентов.

Необходимо провести разработку средств «объясняющей» визуализации результатов прогноза на основе методов LIME и kernel SHAP, а также осуществить применение и анализ результатов этих методов к ошибкам прогноза.

Многие из современных моделей машинного обучения функционально являются черными ящиками. Объяснение причин, лежащих в основе индивидуальных прогнозов, поможет нам больше доверять или не доверять прогнозу или классификатору в целом.

В данной работе рассматриваются 2 средства «объясняющей» визуализации результатов прогноза: SHAP [48, 49, 50], LIME [51, 52, 53]. SHAP основывается на вычислении значения Шепли: это среднее значение предельных вкладов по всем перестановкам. Преимущества SHAP:

1. Глобальная интерпретируемость. Совокупные значения Шепли могут показать, какой вклад каждый предиктор, положительный или отрицательный, вносит в целевую переменную.
2. Локальная интерпретируемость. Каждое наблюдение получает свой собственный набор значений Шепли. Можно объяснить, почему наблюдение получает такое предсказание и вклад предикторов. Локальная интерпретируемость позволяет нам точно определить и сопоставить влияние факторов.
3. Значения SHAP могут быть рассчитаны для любой древовидной модели.

LIME (локально интерпретируемые модельно-агностические объяснения) выполняет следующие шаги:

1. Создание новых образцов и получение прогнозов с использованием исходной модели
2. Взвешивание новых образцов по близости к объясненному экземпляру
3. Построение линейной регрессии для вновь созданных выборок

Основное преимущество LIME перед SHAP – скорость работы. LIME изменяет данные вокруг отдельного прогноза для построения модели, в то время как SHAP должен вычислять все перестановки глобально, чтобы получить локальную точность.

На рисунках 26-33 представлены корректные и некорректные прогнозы для задач классификации 01-234 и 012-34 и визуализация данных прогнозов с помощью моделей SHAP и LIME.

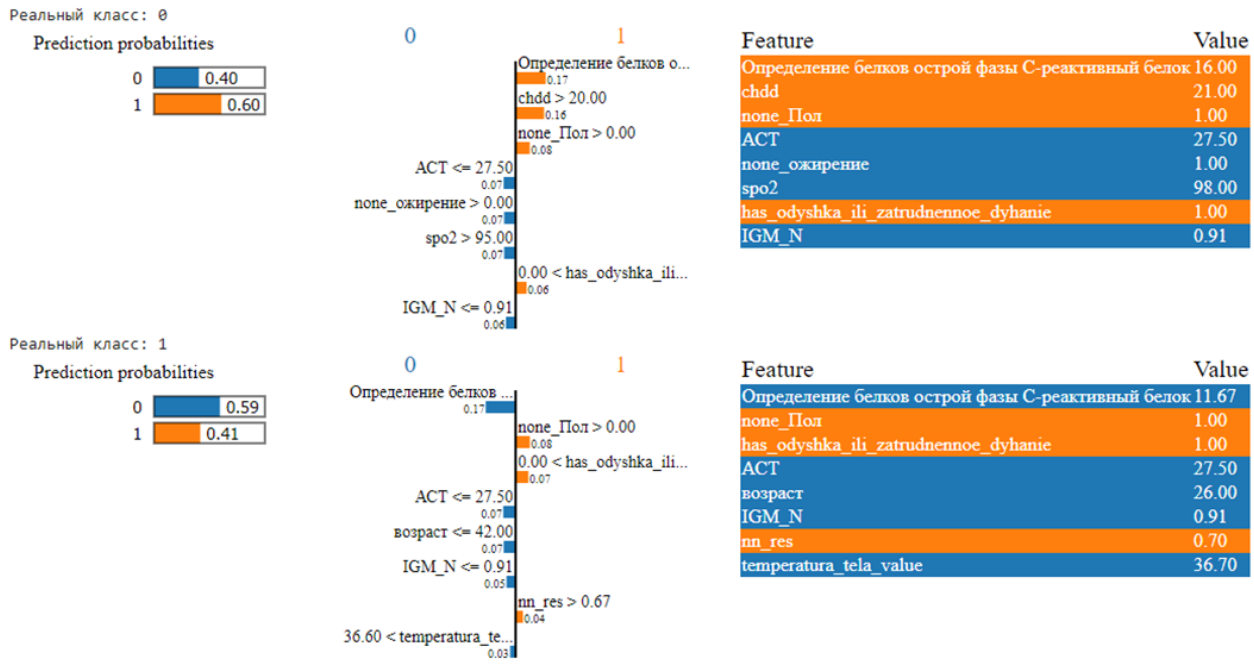

Рисунок 26 — Ошибочная классификация 01-234 LIME

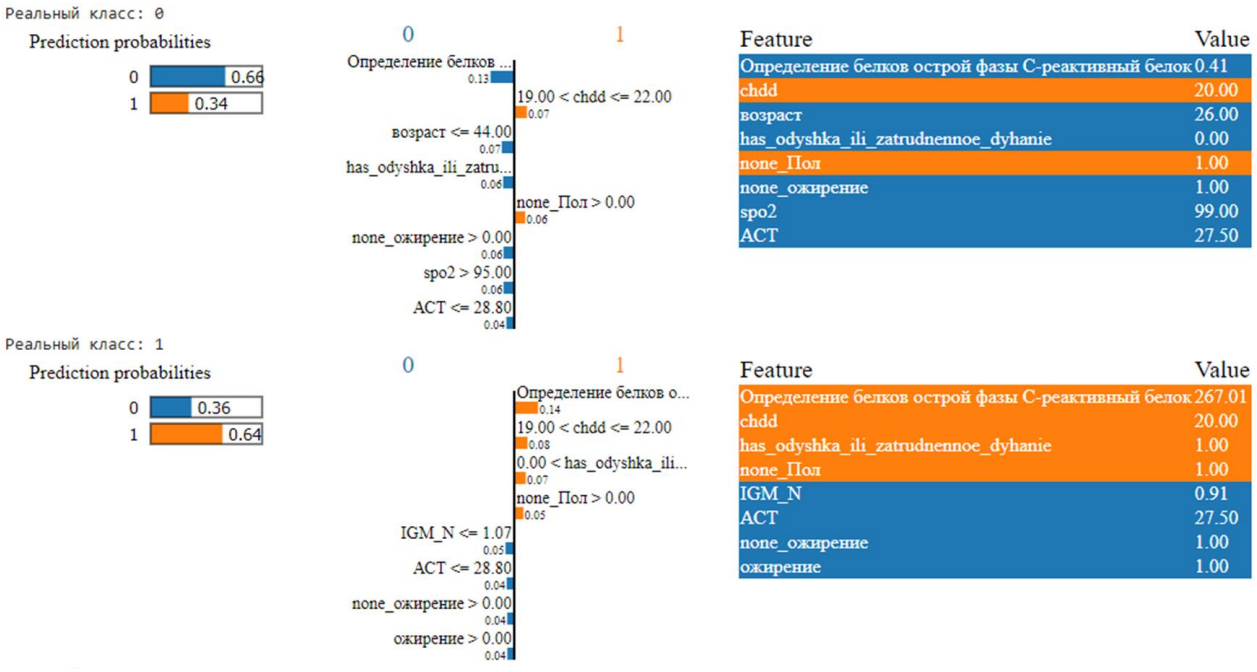

Рисунок 27 — Верная классификация 01-234 LIME

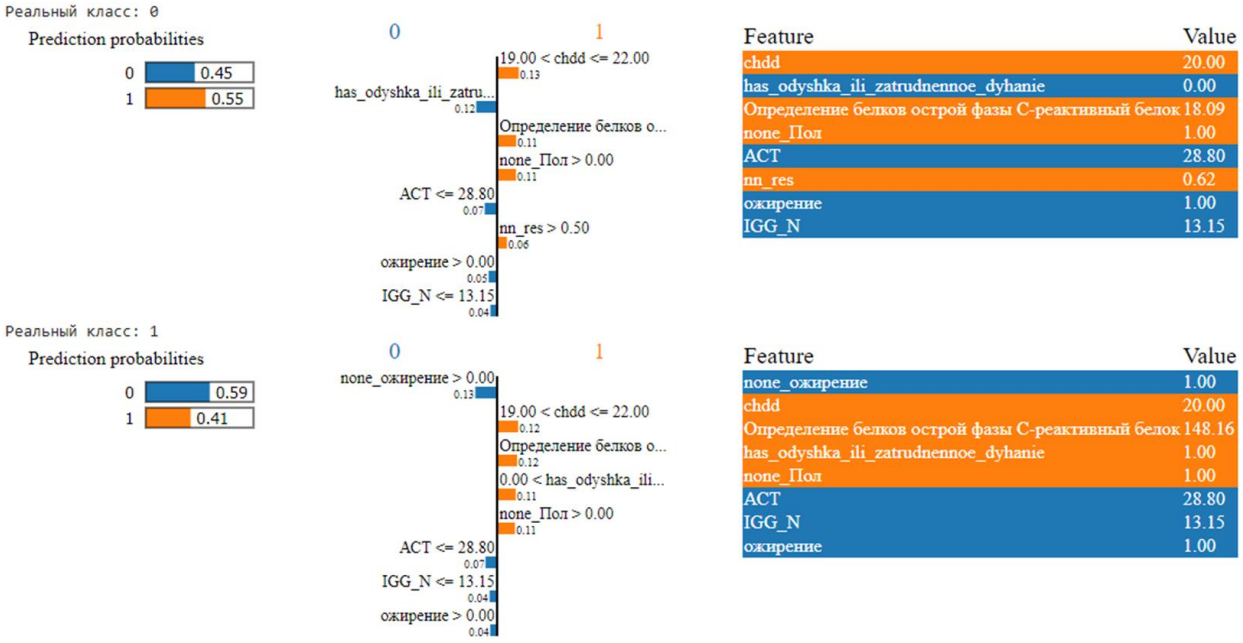

Рисунок 28 — Ошибочная классификация 012-34 LIME

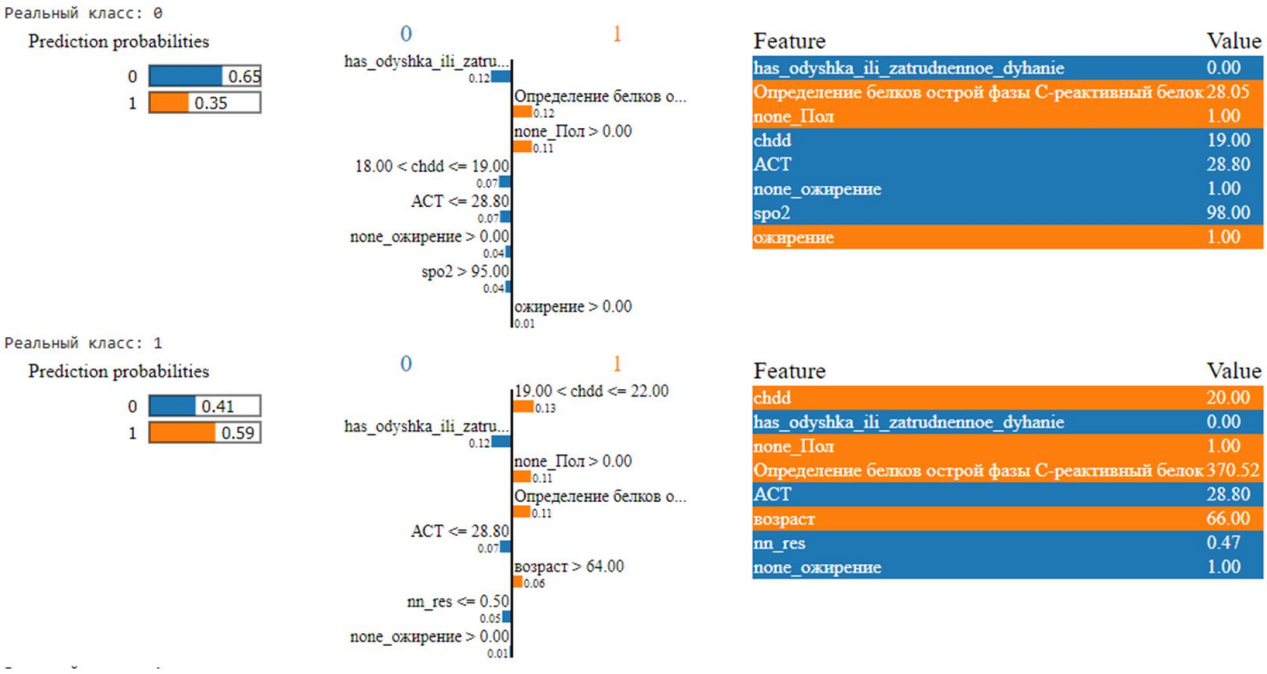

Рисунок 29 — Верная классификация 012-34 LIME

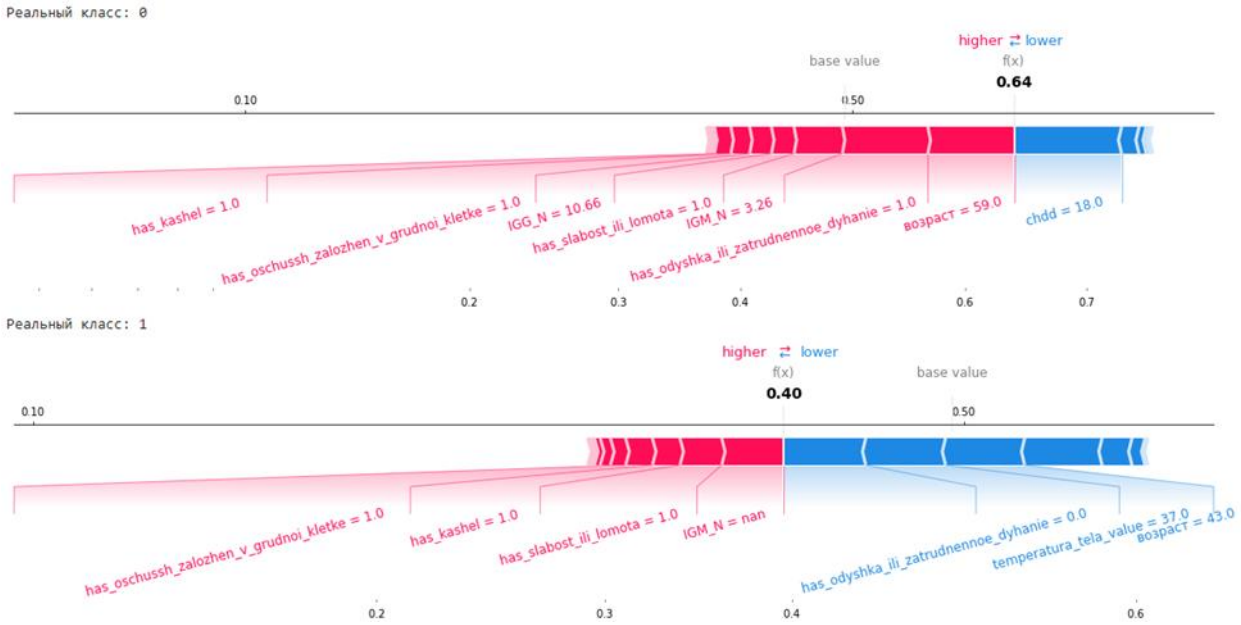

Рисунок 30 — Ошибочная классификация 01-234 SHAP

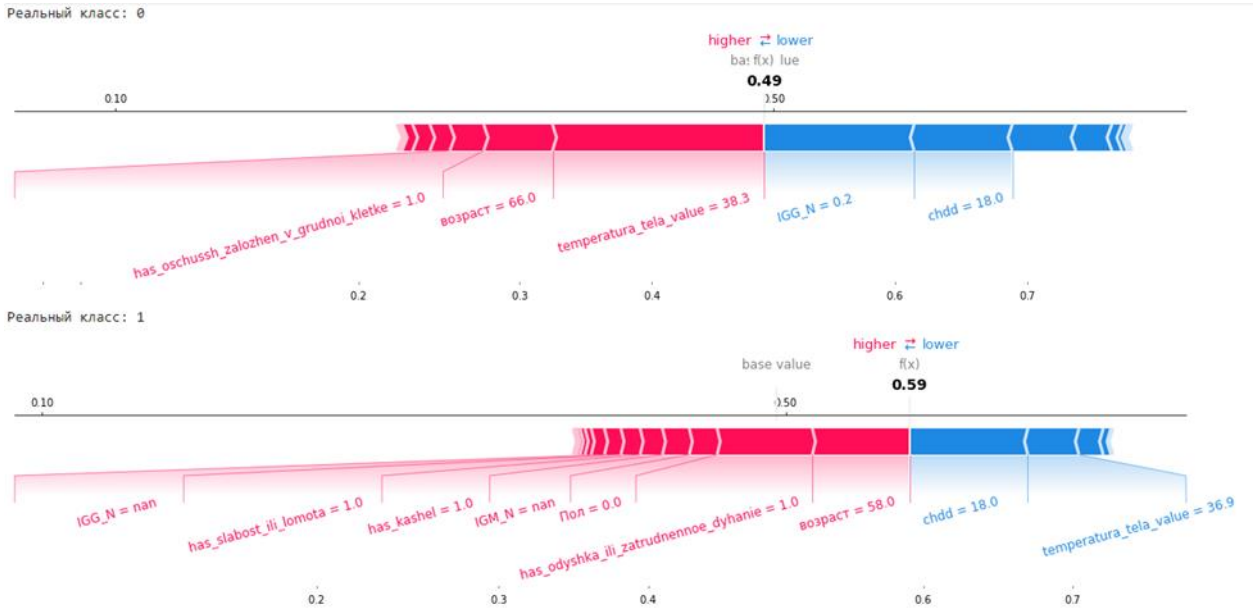

Рисунок 31 — Верная классификация 01-234 SHAP

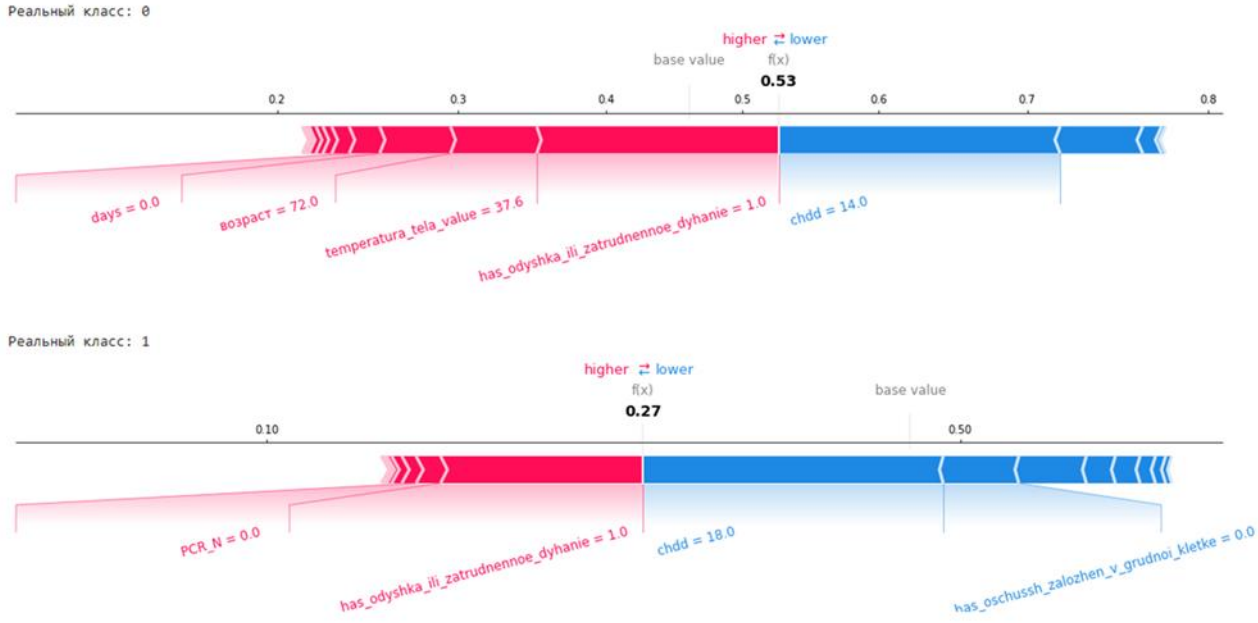

Рисунок 32 — Ошибочная классификация 012-34 SHAP

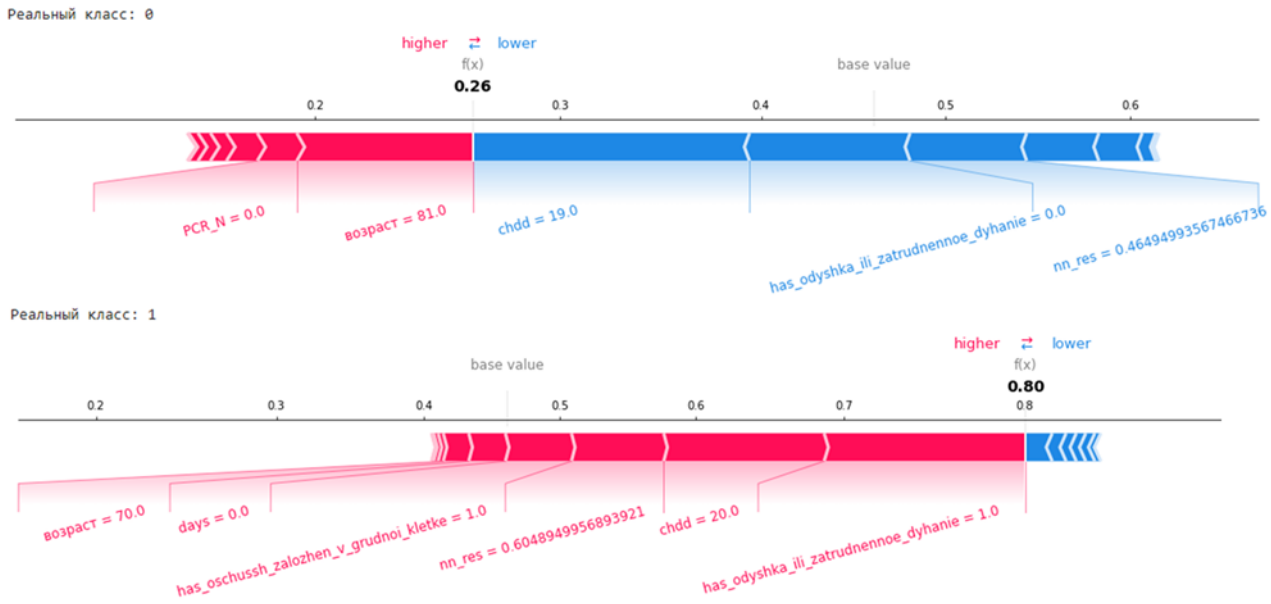

Рисунок 33 — Верная классификация 012-34 SHAP

3.2.12 Выводы

В рамках данного этапа исследований была проведена доработка моделей машинного обучения, используемых в сервисе КТ-калькулятора.

Были предложены алгоритмы дополнительной очистки и нормализации тренировочной выборки с удалением противоречий, а также было проведено расширение признакового пространства, в том числе за счет включения информации о датах проведения анализа и осмотра. Был предложен взаимообратный учет показателя сатурации в оценке степени тяжести с пересчетом по баллам шкалы NEWS.

Был проведен анализ возможности раздельного прогнозирования степени тяжести КТ только по анализу крови с помощью нейросети и только по результатам осмотра с помощью ансамблей деревьев решений для последующего объединения откликов моделей через ансамбль. Также были рассмотрены использование бустинг ансамбля деревьев решений LGBM и ансамбля регуляризованных нейросетей для прогнозирования степени тяжести КТ.

Были предложены подходы для автоматизированного тюнинга и регуляризации моделей (перебор параметров по заданной сетке, система callbacks при обучении), а также подходы для калибровки результата и выбора порогов (калибровка Платта, изотоническая регрессия).

Были предложены методы непротиворечивого сведения откликов двух независимых бинарных моделей, прогнозирующих отдельно вероятность тяжелого поражения (КТ34) и

легкого поражения (КТ01) к многоклассовой вероятностной модели (КТ01, КТ2, КТ34) и к единой шкале оценки тяжести.

Для проверки качества работы предложенных методов была проведена разработка окружения и скриптов (для автоматизированного тестирования сервиса «Калькулятор КТ» через интерфейс REST API), а также средств «объясняющей» визуализации результатов прогноза на основе методов LIME и kernel SHAP.

По результатам экспериментов, наиболее перспективной для внедрения в Калькулятор КТ стала модель LGBM с заполнением пропуском и калибровкой Платта. Данная модель была встроена в тестовую версию Калькулятора КТ. Также высокое качество работы продемонстрировал ансамбль нейросетей.

Процесс отбора наиболее значимых признаков построенных моделей подробно описан в Приложении А данного отчета.

3.3 Выводы

В рамках данного этапа исследований были проведены работы по усовершенствованию сервиса КТ-калькулятор, созданного на предыдущих этапах исследований. В частности, были усовершенствованы используемые алгоритмы построения и применения модели, а также доработана архитектура программы.

Были предложены новые подходы к предобработке данных, а также разработана методика для формирования набора данных для прогнозирования КТ на основе физикальных и клинических признаков. Были предложены алгоритмы дополнительной очистки и нормализации тренировочной выборки с удалением противоречий, а также было проведено расширение признакового пространства, в том числе за счет включения информации о датах проведения анализа и осмотра.

Был предложен взаимообратный учет показателя сатурации в оценке степени тяжести с пересчетом по баллам шкалы NEWS. Был проведен анализ возможности раздельного прогнозирования степени тяжести КТ только по анализу крови с помощью нейросети и только по результатам осмотра с помощью ансамблей деревьев решений для последующего объединения откликов моделей через ансамбль. Также были рассмотрены использование бустинг ансамбля деревьев решений LGBM и ансамбля регуляризованных нейросетей для прогнозирования степени тяжести КТ. Были предложены подходы для автоматизированного тюнинга и регуляризации моделей, а также подходы для калибровки результата и выбора порогов. Были предложены методы непротиворечивого сведения откликов двух независимых бинарных моделей, прогнозирующих отдельно вероятность тяжелого поражения (КТ34) и легкого поражения (КТ01) к многоклассовой вероятностной модели (КТ01, КТ2, КТ34) и к единой шкале оценки тяжести.

По результатам экспериментов, наиболее перспективной для внедрения в Калькулятор КТ стала модель LGBM с заполнением пропусков и калибровкой Платта. Данная модель была встроена в тестовую версию Калькулятора КТ. Также высокое качество работы продемонстрировал ансамбль нейросетей.

4 Заключение

В рамках данного этапа работ были проведены научно-исследовательские работы в области разработки и внедрения методов искусственного интеллекта и анализа больших данных в сфере здравоохранения. Основной целью данного этапа работ являлась доработка функционала сервиса «Калькулятор КТ» – информационной системы для получения экспресс-оценки изменений легочной ткани при COVID-19 без применения компьютерной томографии органов грудной клетки на основе физикальных и лабораторных анализов пациента, созданного на предыдущих этапах исследований.

Решены все поставленные задачи:

1.1 Сформированы проверочные выборки и проведены оценки и доработки моделей-кандидатов на включение в сервис «Калькулятор КТ», разработанный на предыдущих этапах исследований и дорабатываемый на текущем этапе с использованием следующих источников данных:

- Данные о госпитализированных пациентах, собранные в результате экспериментальной работы с сервисом «Калькулятор КТ» в ГБУЗ «ГКБ № 67 им. Л.А. Ворохобова ДЗМ»;
- Данные об амбулаторных пациентах, собранные в результате экспериментальной работы с сервисом «Калькулятор КТ»;
- Данные о пациентах из набора открытых данных изображений КТ легких при COVID-19 Технического университета Хуажонг [2].
- Данные ДИТ от 15.02.21 обо всех клинических и амбулаторных COVID-19 пациентах:
    - общие данные по выжившим и умершим пациентам (включая анамнез) с марта 2020 года до февраля 2021 года;
    - данные о проведенных клинических и биохимических анализах с марта 2020 года до февраля 2021 года;
    - данные из КТ-центров (степень тяжести, а также осмотровые признаки) с марта 2020 года до декабря 2020 года;
    - данные о проведенных тестах ПЦР и ИФА с марта 2020 года до декабря 2020 года;

- данные о взятой сатурации с марта 2020 года до февраля 2021 года.

1.2 Проведена доработка моделей машинного обучения и выбраны лучшие модели для сервиса «Калькулятор КТ» на основе проверки следующих вариантов:

- Раздельное прогнозирование тяжести только по анализу крови с помощью нейросети и только по результатам осмотра с помощью ансамблей деревьев решений (алгоритм машинного обучения «случайный лес») для последующего объединения откликов моделей через ансамбль;
- Использование бустинг ансамбля деревьев решений lgbm (алгоритм машинного обучения «градиентный бустинг») для прогнозирования степени тяжести КТ;
- Использование ансамбля регуляризованных нейросетей для прогнозирования степени тяжести КТ;
- Дополнительная очистка и нормализация тренировочной выборки с удалением противоречий;
- Расширение признакового пространства, в том числе за счет включения информации о датах проведения анализов и осмотра;
- Взаимообратный учет показателя сатурации в оценке степени тяжести с пересчетом по баллам шкалы NEWS;
- Алгоритмы для автоматизированного тьюнинга и регуляризации моделей;
- Исследование различных видов калибровок результата и выбора порогов;
- Методы непротиворечивого сведения откликов двух независимых бинарных моделей, прогнозирующих отдельно вероятность тяжелого поражения (КТ34) и легкого поражения (КТ01) к многоклассовой вероятностной модели (КТ01, КТ2, КТ34) и к единой шкале оценки тяжести;
- Разработка окружения и скриптов для автоматизированного тестирования сервиса «Калькулятор КТ» через интерфейс REST API;
- Разработка средств «объясняющей» визуализации результатов прогноза на основе методов LIME и kernel SHAP, применение и анализ результатов этих методов к ошибкам прогноза.

По результатам работы были опубликованы следующие научные статьи в журналах, индексируемых RSCI:

- 2024 Применение регуляризации при вычислении критериев разбиения в моделях анализа выживаемости. Васильев Ю. А., Петровский М. И., Машечкин И. В. в журнале «Вычислительные методы и программирование». — Т. 25, № 3. — С. 9. DOI: 10.26089/NumMet.v25r328 **(RSCI)**;
- 2024 Разработка библиотеки древовидных моделей анализа выживаемости. Васильев Ю. А. в журнале «Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика». — № 3. — С. 60–72. DOI: 10.55959/MSU/0137-0782-15-2024-47-3-60-72 **(RSCI)**.

Дополнительно, было получено свидетельство о регистрации прав на программное обеспечение:

- 2024 Библиотека методов машинного обучения для построения моделей анализа выживаемости (SOFT). Авторы: Васильев Ю.А., Петровский М.И., Машечкин И.В. #2024681935, 16 сентября.

Также за отчетный период была опубликована научная статья в журнале, индексируемом RSCI, по результатам, полученным на предыдущих этапах данной научно-исследовательской работы:

- 2024 Методы машинного обучения для анализа и моделирования поведения пользователей компьютерных систем. Машечкин И.В., Петровский М.И., Казачук М.А. в журнале Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика, издательство Изд-во Моск. ун-та (М.), том 2024, № 4, с. 160-189. DOI: 10.55959/MSU/0137-0782-15-2024-47-4-160-189 **(RSCI)**.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Das S., Ayus I., Gupta D. A comprehensive review of COVID-19 detection with machine learning and deep learning techniques //Health and Technology. – 2023. – Т. 13. – №. 4. – С. 679-692.
2. ICTCF [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2021. — Режим доступа: http://ictcf.biocuckoo.cn/ — 11.03.2024
3. Solayman S. et al. Automatic COVID-19 prediction using explainable machine learning techniques //International Journal of Cognitive Computing in Engineering. – 2023. – Т. 4. – С. 36-46.
4. Laatifi M. et al. Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME //Scientific Reports. – 2023. – Т. 13. – №. 1. – С. 5481.
5. Ишемическая болезнь сердца [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2024. — Режим доступа: http://мкб11.рф/i20-i25-ишемическая-болезнь-сердца/ — 11.03.2024.
6. Артериальная гипертензия [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2024. — Режим доступа: http://мкб11.рф/i10-i15-болезни-характеризующиеся-повыше/— 11.03.2021.
7. Сахарный диабет [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2021. — Режим доступа: http://мкб11.рф/e10-e14-сахарный-диабет/ — 11.03.2024.
8. Хронические болезни нижних дыхательных путей [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2021. — Режим доступа: http://мкб11.рф/j40-j47-хронические-болезни-нижних-дыхате/ — 11.03.2024.
9. Ожирение и другие виды избыточности питания [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2021. — Режим доступа: http://мкб11.рф/e65-e68-ожирение-и-другие-виды-избыточност/— 11.03.2024.
10. Narkhede S. Understanding auc-roc curve //Towards Data Science. – 2018. – Т. 26. – С. 220-227.
11. Bishop C. M. Pattern recognition and machine learning. – springer, 2006.
12. Ripley B. D. Pattern recognition and neural networks. – Cambridge university press, 2007.
13. Vapnik V. The nature of statistical learning theory. – Springer science & business media, 2013.
14. Breiman L. Random forests //Machine learning. – 2001. – Т. 45. – №. 1. – С. 5-32.
15. Amit Y., Geman D. Shape quantization and recognition with randomized trees //Neural computation. – 1997. – Т. 9. – №. 7. – С. 1545-1588.

16. Geurts P., Ernst D., Wehenkel L. Extremely randomized trees //Machine learning. – 2006. – Т. 63. – №. 1. – С. 3-42.
17. Bauer E., Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants //Machine learning. – 1999. – Т. 36. – №. 1. – С. 105-139.
18. Breiman L. Some infinity theory for predictor ensembles. – Technical Report 579, Statistics Dept. UCB, 2000.
19. Breiman L. Randomisation outputs to increase prediction accuracy. – Technical report, Statistics department, University of California, Berkeley, 1998.
20. Rumelhart D. E., Hinton G. E., Williams R. J. Learning representations by back-propagating errors //nature. – 1986. – Т. 323. – №. 6088. – С. 533-536.
21. Minsky M. L., Papert S. A. Perceptrons: expanded edition. – 1988.
22. Krogh A. What are artificial neural networks? //Nature biotechnology. – 2008. – Т. 26. – №. 2. – С. 195-197.
23. Hunt K. J. et al. Neural networks for control systems—a survey //Automatica. – 1992. – Т. 28. – №. 6. – С. 1083-1112.
24. Bylander T. Learning linear threshold approximations using perceptrons //Neural computation. – 1995. – Т. 7. – №. 2. – С. 370-379.
25. Hastie T., Tibshirani R., Friedman J. Boosting and additive trees //The elements of statistical learning. – Springer, New York, NY, 2009. – С. 337-387.
26. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – С. 785-794.
27. Caruana R., Niculescu-Mizil A. 1.6 A PREDICTING GOOD PROBABILITIES WITH SUPERVISED LEARNING. – 2005.
28. Alzamzami F., Hoda M., El Saddik A. Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation //IEEE Access. – 2020. – Т. 8. – С. 101840-101858.
29. Tutica L. et al. Invoice Deduction Classification Using LGBM Prediction Model //Advances in Electronics, Communication and Computing. – Springer, Singapore, 2021. – С. 127-137.
30. Gul A. et al. Ensemble of a subset of k NN classifiers //Advances in data analysis and classification. – 2018. – Т. 12. – №. 4. – С. 827-840.
31. Rosen B. E. Ensemble learning using decorrelated neural networks //Connection science. – 1996. – Т. 8. – №. 3-4. – С. 373-384.
32. Yadav D. C., Pal S. Prediction of thyroid disease using decision tree ensemble method //Human-Intelligent Systems Integration. – 2020. – Т. 2. – №. 1. – С. 89-95.

33. Dada O. A. et al. The relationship between red blood cell distribution width and blood pressure in patients with type 2 diabetes mellitus in Lagos, Nigeria //Journal of blood medicine. – 2014. – Т. 5. – С. 185.
34. Khazaal M. S., Hamdan F. B., Al- Mayah Q. S. Association of BCR/ABL transcript variants with different blood parameters and demographic features in Iraqi chronic myeloid leukemia patients //Molecular genetics & genomic medicine. – 2019. – Т. 7. – №. 8. – С. e809.
35. Palareti G. et al. D-dimer testing to determine the duration of anticoagulation therapy //New England Journal of Medicine. – 2006. – Т. 355. – №. 17. – С. 1780-1789.
36. Протокол оценки тяжести состояния пациента (NEWS) [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2021. — Режим доступа: https://euat.ru/upload/images/1586077839.pdf — 11.03.2021
37. Ndiaye E. et al. Safe grid search with optimal complexity //International Conference on Machine Learning. – PMLR, 2019. – С. 4771-4780.
38. DeGroot M. H., Fienberg S. E. The comparison and evaluation of forecasters //Journal of the Royal Statistical Society: Series D (The Statistician). – 1983. – Т. 32. – №. 1-2. – С. 12-22.
39. Niculescu-Mizil A., Caruana R. Obtaining Calibrated Probabilities from Boosting //UAI. – 2005. – Т. 5. – С. 413-20.
40. Platt J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods //Advances in large margin classifiers. – 1999. – Т. 10. – №. 3. – С. 61-74.
41. Kull M. et al. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration //Electronic Journal of Statistics. – 2017. – Т. 11. – №. 2. – С. 5052-5080.
42. Stephens M. A. Use of the Kolmogorov–Smirnov, Cramer–Von Mises and related statistics without extensive tables //Journal of the Royal Statistical Society: Series B (Methodological). – 1970. – Т. 32. – №. 1. – С. 115-122.
43. Daniel W. W. Kolmogorov–Smirnov one-sample test //Applied nonparametric statistics. – 1990. – Т. 2.
44. Marsaglia G. et al. Evaluating Kolmogorov's distribution //Journal of statistical software. – 2003. – Т. 8. – №. 18. – С. 1-4.
45. Shorack G. R., Wellner J. A. Empirical processes with applications to statistics. – Society for Industrial and Applied Mathematics, 2009.
46. Zhou W. et al. REST API design patterns for SDN northbound API //2014 28th international conference on advanced information networking and applications workshops. – IEEE, 2014. – С. 358-365.

47. Masse M. REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces. – " O'Reilly Media, Inc.", 2011.
48. Using SHAP and LIME [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2021. — Режим доступа: https://blog.dominodatalab.com/shap-lime-python-libraries-part-2-using-shap-lime/ — 11.03.2021.
49. Lundberg S. M., Erion G. G., Lee S. I. Consistent individualized feature attribution for tree ensembles //arXiv preprint arXiv:1802.03888. – 2018.
50. Ribeiro M. T., Singh S., Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier //Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. – 2016. – С. 1135-1144.
51. LIME - Local Interpretable Model-Agnostic Explanations [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2021. — Режим доступа: https://homes.cs.washington.edu/~marcotcr/blog/lime/ — 11.03.2021.
52. Craven M. W., Shavlik J. W. Extracting tree-structured representations of trained networks //Advances in neural information processing systems. – 1996. – С. 24-30.
53. Dzindolet M. T. et al. The role of trust in automation reliance //International journal of human-computer studies. – 2003. – Т. 58. – №. 6. – С. 697-718
54. Anderson C. Docker [software engineering] //Ieee Software. – 2015. – Т. 32. – №. 3. – С. 102-c3.

Приложение А. Процедуры отбора значимых признаков

Были рассмотрены следующие процедуры отбора значимых признаков:

- Random Forest;
- Lasso;
- Stepwise Backward.

A.1 Random Forest

Случайные леса, используемые для классификации объектов, могут естественным образом быть использованы для оценки важности признаков. Существует несколько способов определения важностей признаков. Во-первых, может быть рассчитана «важность Джини», определяемая как общее уменьшение примесей в узле (взвешенное по вероятности достижения этого узла, которая аппроксимируется долей выборок, достигших этого узла), усредненное по всем деревья ансамбля.

Во втором подходе во время построения модели для каждой подвыборки обучающей выборки рассчитывается out-of-bag ошибка (усредненная оценка алгоритмов на тех данных, на которых они не обучались). Для каждой подвыборки такая ошибка усредняется по всему случайному лесу. Чтобы оценить важность конкретного признака, его значения перемешиваются для всей обучающей выборки и out-of-bag ошибка считается снова. Важность параметра подсчитывается путем усреднения по всем деревьям разности показателей out-of-bag ошибок до и после перемешивания значений.

A.2 Lasso

В методе LASSO коэффициенты линейной модели $y = X\beta + \varepsilon$ находятся из решения задачи минимизации среднеквадратичной ошибки. При этом в функцию потерь добавляется штраф на величину коэффициентов: $loss(\beta, \lambda) = \| y - X\beta \| + \lambda \sum_i |\beta i|$, где λ — параметр регуляризации, который задаёт баланс между качеством подгонки модели и её сложностью. Для того, чтобы получить аналитическое решение этой задачи, необходимо для каждого коэффициента минимизировать функцию потерь.

Величина полученных коэффициентов пропорциональна важности соответствующих переменных в модели, а для переменных, которые дают наименьший вклад в устранение ошибки, коэффициенты станут нулевыми.

A.3 Stepwise Backward

Метод основан на последовательном исключении наименее значимого параметра. На каждой итерации каждый параметр проходит проверку на значимость и удаляется параметр, имеющий наибольшее p-value. Для каждой модели рассчитывается индекс Акаике (AIC), модель с наименьшим AIC объявляется лучшей.

A.4 Таблица результатов

На основе описанных процедур отбора признаков был произведен расчет важностных характеристик признаков. Были рассмотрены 7 основных важностных характеристик:

- RF;
- GDB;
- Odds Lasso;
- Odds OLS;
- P-value OLS;
- Odds backward;
- P-value backward.

Понятие p-value тесно связано с понятием нулевой гипотезы. Нулевая гипотеза – это принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями A и B. P-value – это вероятность ошибки при отклонении нулевой гипотезы.

Решение о принятии или отклонении нулевой гипотезы принимается в результате сравнения p-value с выбранным уровнем значимости. Если оно превышает указанный уровень значимости, то для отклонения нулевой гипотезы (принятия альтернативной) нет достаточных оснований. Таким образом, чем ниже p-value, тем выше вероятность в том, что наблюдаемые события A и B связаны между собой. Зачастую, в медицинских исследованиях выбирается значение уровня значимости, равное 0.05 (5%), однако часто используются значения уровня значимости, равные 0.01 или 0.005.

Заметим, что с помощью p-value мы можем оценить значимость признаков обучающей выборки, приняв за событие B факт использования данного признака в обучающей модели, а за событие A – факт верной классификации.

Термин отношение шансов (ОШ, OR или Odds ratio) определяет силу взаимосвязи между двумя событиями (признаками) A и B в пределах одной и той же выборки.

Отношение шансов вычисляется по формуле: ОШ = (A+/A-) / (B+/B-), где

- A+ – вероятность того, что событие A произойдет;
- A- – вероятность того, что событие A не произойдет;
- B+ – вероятность того, что событие B произойдет;
- B- – вероятность того, что событие B не произойдет.

Если odds > 1, то появление B повышает шансы наличия A (по отношению к отсутствию B). Если odds < 1, то наличие одного события (признака) уменьшает шансы другого события (признака). Если odds = 1, то события независимы.

Заметим, что с помощью отношения шансов мы можем оценить значимость признаков обучающей выборки, приняв за событие B факт использования данного признака в обучающей модели, а за событие A – факт верной классификации.

Далее представлены 4 таблицы результатов. Таблицы 1 и 3 содержат сведения о 20 наиболее значимых признаках относительно каждой процедуры отбора признаков. Признаки расположены в порядке убывания значимости.

Признаки с префиксом «n» отражают наличие или отсутствие значения в исходном признаке. Признаки с постфиксом «d» (IGG, IGM) отражают выход за границы порога (для IGG пороговое значение – 10, для IGM пороговое значение – 2).

Таблицы 2 и 4 содержат сведения о 20 наиболее значимых признаках относительно всех процедур отбора признаков. Для расчета общей значимости признаков, каждому признаку был сопоставлен ранговый номер в процедурах отбора признаков. На основе полученных рангов был рассчитан как средний ранг каждого признака (mean rank), так и количество попаданий признака в топ 20 по каждой процедуре (top rank). Все признаки были отсортированы по возрастанию среднего ранга и в таблицу определены первые 20 признаков.

Таблица 1 — Классификация КТ 01-234: топ 20 признаков каждой из процедур

RF	**Gdb**	**odds lasso**	**p-value ols**	**odds ols**	**odds backward**	**p-value backward**
Сатурация	Сатурация	C-	C-	C-	C-	C-

RF	Gdb	odds lasso	p-value ols	odds ols	odds backward	p-value backward
		реактивный белок	реактивный белок	реактивный белок	реактивный белок	реактивный белок
С-реактивный белок	С-реактивный белок	Одышка	Одышка	Одышка	Одышка	Одышка
ЧДД	ЧДД	возраст	возраст	возраст	НСТ	возраст
АСТ	IGM	НСТ	Температура тела	НСТ	возраст	Температура тела
Температура тела	Температура тела	Температура тела	ЧДД	Температура тела	Температура тела	ЧДД
Возраст	IGG	nIGG	СОЭ по Вестергрену	ЧДД	ЧДД	СОЭ по Вестергрену
LYM	возраст	ЧДД	Сатурация	nАбс. кол-во эозинофилов	nАбс. кол-во эозинофилов	Сатурация
IGM	PLT	nIGMd	НСТ	nGRA	nGRA	НСТ
Одышка	nСатурация	nАбс. кол-во эозинофилов	nIGMd	СОЭ по Вестергрену	СОЭ по Вестергрену	nLYM
IGG	RDW	СОЭ по Вестергрену	nIGGd	nLYM	Сатурация	nIGMd
АЛТ	MON	nGRA	nIGG	Сатурация	IGMd	nIGGd
PLT	ACT	nLYM	nIGM	IGMd	nLYM	nIGM
WBC	Количество дней до ОАК	Сатурация	nСатурация	IGGd	IGGd	nIGG

RF	Gdb	odds lasso	p-value ols	odds ols	odds backward	p-value backward
RDW	Определение креатинина	IGMd	Заложенность	nIGMd	nIGGd	Заложенность
Определение креатинина	Количество дней до БАК	IGGd	IGMd	nIGGd	nIGMd	nСатурация
MON	HCT	nPDW	Слабость	nPDW	nIGM	IGMd
HCT	WBC	Заложенность	Пол	nIGM	nIGG	nАбс. кол-во эозинофилов
RBC	СОЭ по Вестергрену	nСатурация	IGGd	nIGG	nАбс. кол-во нейтрофилов	Пол
HGB	Исследование ферритина	nАбс. кол-во нейтрофилов	nPDW	nАбс. кол-во нейтрофилов	none_PDW	Слабость
MPV	АЛТ	Пол	nАбс. кол-во эозинофилов	Заложенность	Заложенность	IGGd

Таблица 2 — Классификация КТ 01-234: топ 20 признаков с учетом рангов

Признак	RF	gdb	odds lasso	p-value ols	odds ols	odds backward	p-value backward	top rank	mean rank
С-реактивный белок	1	1	0	0	0	0	0	7	0,285714

Признак	RF	gdb	odds lasso	p-value ols	odds ols	odds backward	p-value backward	top rank	mean rank
Возраст	5	6	2	2	2	3	2	7	3,142857
Температура тела	4	4	4	3	4	4	3	7	3,714286
ЧДД	2	2	6	4	5	5	4	7	4
Сатурация	0	0	12	6	10	9	6	7	6,142857
Одышка	8	34	1	1	1	1	1	6	6,714286
НСТ	16	15	3	7	3	2	7	7	7,571429
СОЭ по Вестергрену	21	17	9	5	8	8	5	6	10,42857
nСатурация	29	8	17	12	20	20	14	4	17,14286
IGM	7	3	24	26	26	26	27	2	19,85714
Заложенность	28	33	16	13	19	19	13	5	20,14286
nIGG	47	35	5	10	17	16	12	5	20,28571
nLYM	60	24	11	23	9	11	8	4	20,85714
ACT	3	11	25	30	27	27	30	2	21,85714
nPDW	43	25	15	18	15	18	20	4	22
IGMd	31	62	13	14	11	10	15	5	22,28571
Количество дней до БАК	24	14	23	25	25	25	26	1	23,14286
Пол	38	32	19	16	21	21	17	3	23,42857
Исследование ферритина	35	18	21	22	23	23	23	1	23,57143
IGGd	32	61	14	17	12	12	19	5	23,85714

Таблица 3 — Классификация КТ 012-34: топ 20 признаков каждой из процедур

RF	gdb	odds lasso	p-value ols	odds ols	odds backward	p-value backward
chdd	Сатурация	Одышка	Одышка	Одышка	Одышка	Одышка

RF	**gdb**	**odds lasso**	**p-value ols**	**odds ols**	**odds backward**	**p-value backward**
Сатурация	С-реактивный белок	С-реактивный белок	С-реактивный белок	С-реактивный белок	С-реактивный белок	С-реактивный белок
С-реактивный белок	chdd	chdd	chdd	chdd	chdd	chdd
Одышка	IGG	Сатурация	Сатурация	Сатурация	Сатурация	Сатурация
ACT	RDW	nАбс. кол-во нейтрофилов	возраст	nАбс. кол-во нейтрофилов	nАбс. кол-во нейтрофилов	Заложенность
IGG	ACT	Заложенность	Заложенность	Заложенность	Заложенность	возраст
LYM	Температура тела	возраст	Количество дней до БАК	возраст	возраст	ACT
IGM	возраст	Количество дней до БАК	nвозраст	Количество дней до БАК	ACT	Количество дней до БАК
Заложенность	PLT	ACT	IGG	ACT	Количество дней до БАК	IGG
Температура тела	АЛТ	nIGG	Пол	nАбс. кол-во базофилов	IGG	Пол
WBC	Определение креатинина	IGG	PCR	IGG	nАбс. кол-во базофилов	nIGG
Абс. кол-во нейтрофил	MON	nАбс. кол-во	ACT	Количество дней до	Количество дней до	nIGM

RF	gdb	odds lasso	p-value ols	odds ols	odds backward	p-value backward
ов		базофилов		ОАК	ОАК	
RDW	LYM	Количество дней до ОАК	nАбс. кол-во нейтрофилов	Пол	IGGd	PCR
Возраст	IGM	Пол	Температура тела	НСТ	Пол	nMON
АЛТ	Абс. кол-во нейтрофилов	nАбс. кол-во эозинофилов	Количество дней до ОАК	nАбс. кол-во эозинофилов	nАбс. кол-во эозинофилов	nАбс. кол-во нейтрофилов
PLT	WBC	НСТ	Заложенность	PCR	PCR	IGGd
Определение креатинина	nСатурация	PCR	СОЭ по Вестергрену	Температура тела	nMON	Температура тела
RBC	Количество дней до БАК	Температура тела	nIGG	IGGd	НСТ	Количество дней до ОАК
MON	PDW	IGGd	nIGM	nСОЭ_по Вестергрену	Температура тела	nСОЭ_по Вестергрену
НСТ	НСТ	nСОЭ по Вестергрену	nКоличество дней до БАК	nIGG	nСОЭ_по Вестергрену	СОЭ по Вестергрену

Таблица 4 — Классификация КТ 012-34: топ 20 признаков с учетом рангов

Признак	RF	gdb	odds lasso	p-value ols	odds ols	odds backward	p-value backward	top rank	mean rank
С-реактивный	2	1	1	1	1	1	1	7	1,142857

Признак	**RF**	**gdb**	**odds lasso**	**p-value ols**	**odds ols**	**odds backward**	**p-value backward**	**top rank**	**mean rank**
белок									
ЧДД	0	2	2	2	2	2	2	7	1,714286
Сатурация	1	0	3	3	3	3	3	7	2,285714
Одышка	3	24	0	0	0	0	0	6	3,857143
Возраст	13	7	6	4	6	6	5	7	6,714286
АСТ	4	5	8	11	8	7	6	7	7
IGG	5	3	10	8	10	9	8	7	7,571429
Заложенность	8	29	5	5	5	5	4	6	8,714286
Количество дней до БАК	22	17	7	6	7	8	7	6	10,57143
Температура тела	9	6	17	13	16	18	16	7	13,57143
Количество дней до ОАК	24	21	12	14	11	11	17	5	15,71429
Пол	37	34	13	9	12	13	9	5	18,14286
НСТ	19	19	15	31	13	17	22	5	19,42857
PCR	36	32	16	10	15	15	12	5	19,42857
RDW	12	4	21	24	23	29	27	2	20
СОЭ по Вестергрену	23	26	20	16	21	21	19	2	20,85714
nАбсолютное количество нейтрофилов	51	57	4	12	4	4	14	5	20,85714
IGM	7	13	31	35	31	22	21	2	22,85714
nIGG	35	44	9	17	19	26	10	4	22,85714
IGGd	40	38	18	21	17	12	15	4	23