УДК 51-76 DOI: 10.26102/2310-6018/2021.33.2.028

Моделирование биологического возраста пациентов на основе их функциональных показателей

О.В. Лимановская[1,2], И.В. Гаврилов[2], В.Н. Мещанинов[2], Д.Л. Щербаков[2], Е.Н. Колос[1]

[1]ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина», Екатеринбург, Российская Федерация

[2]ФГБОУ ВО «Уральский государственный медицинский университет Министерства здравоохранения Российской Федерации», Екатеринбург, Российская Федерация

Резюме: Процесс старения является сложным многофакторным явлением, на который оказывает влияние, как внешние факторы – климатические, экономические и политические условия, так и индивидуальные особенности организма. В связи с этим моделирование данного процесса является нетривиальной задачей, требующего разностороннего подхода для ее решения. Анализ литературы показывает, что при моделировании темпов старения используются как концептуальные [1-4] модели, дающие представления как в принципе оценивать процесс старения, так и более конкретизированные расчетные модели [5-9], дающие возможность прогнозирования темпов старения. При построении расчетных моделей возникает противоречие между полнотой модели и возможностью ее использования для прогнозирования. Так модели, хорошо показывающие все взаимосвязи в процессе старения [7], построенные как правило, на графах, сложны в применении их к численной оценке темпа старения, хотя некоторые из них дают возможность построения индивидуальных траекторий старения [8-9]. В то же время, модели, имеющие сильный численный аппарат оценки темпа старения [5-6], как правило, заострены для решения узкой задачи и не охватывают всей сложности процесса старения. В такой ситуации использование методов машинного обучения в расчетных моделях оценки темпов старения является очень перспективным направлением [10-15], поскольку его применение позволяет учесть все многообразие факторов процесса старения, не вникая в сущность самого процесса. В данной работе методами машинного обучения проведен анализ корреляции функциональных показателей пациентов с их календарным возрастом и построению моделей прогнозирования биологического возраста пациентов. Анализ данных проводился с помощью авторских разработок на языке Python в среде Anaconda. Для анализа использовались 10 функциональных показателей 1185 пациентов из базы данных клинического областного психико-неврологического госпиталя ветеранов войны. Анализ данных показал наличие статически значимой корреляции используемых показателей с календарным возрастом пациентов. В работе построены 5 моделей регрессии с помощью различных инструментов библиотеки sklern языка Python (пакетный градиентный спуск, стохастический градиентный спуск, гребневая регрессия, гребневая регрессия с Байесовским отбором, метод опорных векторов), а также использовались композиции алгоритмов из решающих деревьев (случайный лес и бустинг). Для улучшения качества модели применялись отбор признаков (add-dell) и поиск и удаление выбросов методом опорных векторов, изолирующего леса и методом ближайших соседей. Все полученные модели адекватны (проверка критерием Фишера), но наибольшую точность ($R^2 = 0,75$) показала модель композиции случайного леса на полном наборе признаков после удаления аномалий методом опорных векторов. Результаты моделирования по линейным моделям показали, что наибольшие веса в модели имеют 3 функциональных показателя – аккомодация, жизненная емкость легких и острота слуха.

Ключевые слова: задача регрессии, отбор признаков, поиск и удаление аномалий, машинное обучение, биологический возраст

Для цитирования: Лимановская О.В., Гаврилов И.В., Мещанинов В.Н., Щербаков Д.Л., Колос Е.Н. Моделирование биологического возраста пациентов на основе их функциональных показателей.

Моделирование, оптимизация и информационные технологии. 2021;9(2). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=966 DOI: 10.26102/2310-6018/2021.33.2.028

Modeling the biological age of the patients based on their functional indicators

O.V. Limanovskaya[1,2], I.V. Gavrilov[2], V.N. Meshchaninov[2], D.L. Shcherbakov[2], E.N. Kolos[1]

[1] FSAEI HE «UrFU named after the first President of Russia B.N.Yeltsin», Ekaterinburg, Russian Federation

[2]FSBEI HE «USMU of the Ministry of Health of the Russian Federation», Ekaterinburg, Russian Federation

Abstract: The aging process is a complex multifactorial phenomenon. It is influenced by both external factors - climatic, economic, and political conditions - and individual characteristics of the body. In this regard, modeling this process is a non-trivial task that requires a versatile approach to solve. The literature analysis shows that when modeling the rate of aging, both conceptual [1-4] models are used, which give an idea of how to assess the aging process in principle, and more specific computational models [5-9], which make it possible to predict the rate of aging. When constructing computational models, there is a contradiction between the completeness of the model and the possibility of using it for forecasting. Thus, models that show all the relationships in the aging process well [7], which are usually constructed on graphs, are hard to apply to the numerical estimation of the aging rate, although several of them make possible individual aging tracing [8-9]. At the same time, models that have a powerful numerical apparatus for estimating the rate of aging [5-6], as a rule, are sharpened to solve a narrow task and do not cover the entire complexity of the aging process. In such a situation, the use of machine learning methods in computational models for estimating the rate of aging is an advanced research direction [10-15], since its application allows us to take into consideration all the variety of factors of the aging process, without delving into the essence of the process itself. In this paper, machine learning methods are used to analyze the correlation of functional indicators of patients with their calendar age and to build models for predicting the biological age of patients. The data analysis was carried out with the help of the author's developments in the Python language in the Anaconda environment. Ten functional indicators of 1185 patients from the clinical regional psycho-neurological hospital of war veterans database were used for the analysis. The research revealed a statically significant correlation of the indicators used with the calendar age of the patients. Five regression models were constructed using various tools of the Python skin library: Batch Gradient Descent, Stochastic Gradient Descent, ridge regression, ridge regression with Bayesian selection, the support vector machine method, and algorithm compositions from decision trees (random forest and boosting) were used. To improve the quality of the model, we used feature selection (add-dell) and outlier search and removal using the reference vector method, the isolating forest method, and the nearest neighbor method. All the models obtained are adequate (verification by the Fisher criterion), but the most accurate (R2 = 0.75) showed the model of the composition of a random forest on the entire range of features after the anomalies removal by the support vector machine. The modeling outcomes using linear models showed that the highest weights in the model have three functional indicators – accommodation, the vital capacity of the lungs, and hearing acuity.

Keywords: regression problem, feature selection, finding and removing anomalies, machine learning, biological age

For citation: Limanovskaya O.V., Gavrilov I.V., Meshchaninov V.N., Shcherbakov D.L., Kolos E.N. Modeling the biological age of patients based on their functional indicators. *Modeling, Optimization and Information Technology*. 2021;9(2). Available from: https://moitvivt.ru/ru/journal/pdf?id=966 DOI: 10.26102/2310-6018/2021.33.2.028 (In Russ).

Введение

Старения является комплексным процессом, затрагивающим практически все органы и функции организма. Поскольку в этот процесс вовлечено множество взаимосвязанных показателей, его моделирование представляется сложной многофакторной задачей, которую можно свести к задаче определения темпов старения. В работе [16] дан подробный анализ различных подходов по определению темпов старения и показано, что феноменологические подходы, основанные на использовании сематических сетей [7-8] и подходы, основанные на определении скорости накопления дегенеративных изменений в организме [5-6] позволяют оценить темп старения только качественно, но не дают количественной оценки. Для количественной оценки темпов старения используются комплексные показатели темпов старения, такие как биологический возраст или FI индекс, отображающий суммарное количество накопленных повреждений в организме.

Биологический возраст отображает реальный возраст пациента на основе его функциональных, психологических, гематологических, генетических показателей. Использование методов машинного обучения для определения биовозраста имеет свои преимущества за счет того, что позволяет использовать большое число факторов [17] и строить модели, учитывающие большое разнообразие показателей, влияющих на биовозраст. Так в работе [11] с помощью искусственных нейронных сетей были построены прогностические модели предсказания биовозраста пациентов на основе их гематологических данных. Наилучшая модель предсказывает возраст с ошибкой 6,07 лет на 10 летнем промежутке. Кроме того, методами машинного обучения можно выявить корреляции между факторами и биовозрастом, не имея какой-либо гипотезы о сути этой взаимосвязи [17]. В работе [11] проведено прогнозирование биовозраста и выявлены предикторы старения методами машинного обучения и нейронных сетей. В работе использовалась база данных из 1 563 пациентов. В своих моделях авторы использовали как полносвязные глубокие нейронные сети, так и методы машинного обучения (градиентный бустинг, случайный лес, решающие деревья, линейную регрессию, метод ближайших соседей и метод опорных векторов). Результаты показали, что модели с использованием полносвязных глубоких нейронных сетей показывают более высокую точность ($R^2 < 0,72$), чем методы машинного обучения ($R^2 > 0,8$). Аналогичная попытка прогнозирования биовозраста на основе биохимических показателей была предпринята в работе [19], но полученные модели показали очень низкую точность (наилучшая модель дала точность $R^2=0,59$). Более практичным с клинической точки зрения, является использование не инвазивных методов контроля состояния пациента, поэтому возможность прогнозирования темпа старения на таких показателях является очень актуальной темой. В работе [18] авторам удалось выявить биомаркеры старения и построить прогностические многофакторные модели на основе данных артериального индекса 303 пациентов, среди которых были пациенты, страдающие диабетом второго типа. Прогностические модели биовозраста были получены отдельно для мужчин и женщин. Несмотря на то, что точность моделей получилась не очень высокой (R2 не превышает 0,7), отклонения расчетного возраста пациентов с диабетом от календарного было выше, чем для здоровых пациентов. Такие результаты показывают, что полученные модели чувствительны к наличию заболевания повышающего вероятность смерти у пациента. Успешное применение методов машинного обучения для оценки биовозраста и выделению предикторов старения из клинических данных вдохновила нас к построению прогностических моделей определения биовозраста на основе функциональных показателей пациентов.

Данная работа посвящена построению прогностических моделей методами машинного обучения для определения биовозраста пациентов на основе их функциональных показателей и выделения биомаркеров старения из анализируемого набора показателей.

Материалы

Данные получены из медицинской организации ГАУЗ СО «СОКП Госпиталь для ветеранов войн» за 1995 – 2014 гг. в объеме 6440 записей данных пациентов в возрасте от 15 до 93 лет. Из полученных данных были выделены функциональные показатели пациентов (10 показателей), впервые пришедших на обследование. Была проведена предварительная обработка информации, заключающаяся в удалении пропущенных значений. Итоговая выборка содержала 10 числовых признаков и 1185 записей. Анализировался 10 показателей из предоставленного набора функциональных параметров:

1. АДС – артериальное давление систолическое в мм.рт.ст.,
2. АДД – артериальное давление диастолическое в мм.рт.ст.,
3. АДП – разность между систолическим и диастолическим давлением в мм.рт.ст.,
4. ЗДВдох – задержка дыхания на вдохе в секундах,
5. ЗДВыдох – задержка дыхания на выдохе в секундах,
6. ЖЕЛ – жизненная емкость легких в мл,
7. масса – масса тела в кг,
8. аккомодация в диоптриях,
9. острота слуха в бел,
10. статическая балансировка в секундах.

Методы

Выявление коллинеарных признаков. Признаки считаются коллинеарными, если имеют корреляцию между собой сильнее, чем с целевой переменной. Наличие коллинеарных признаков ухудшает качество модели, способствует ее переобучению и понижает скорость обучения модели. Также наличие сильной корреляции между признаками затрудняет задачу отбора признаков, поскольку алгоритмы отбора будут ошибаться при выборе признака. Для определения наличия и силы корреляции между признаками используются множество методов корреляционного анализа, в зависимости от типа признаков. Наличие и силу корреляции между вещественными признаками можно определить с помощью коэффициента корреляции Пирсона, коэффициента корреляции Спирмена и линейного коэффициента корреляции [20]. Все коэффициенты изменяются от -1 до 1. Значение -1 свидетельствует о наличие обратной линейной связи между признаками, значение 1 – о наличии полной линейной связи между признаками и возможности замены одного признака другим. Значения коэффициента корреляции близкого к 0 свидетельствует о незначительной взаимосвязи между признаками или об ее отсутствии. Коэффициент корреляции Пирсона и коэффициент линейной корреляции определяет силу линейной взаимосвязи двух вещественных показателей. Ранговый коэффициент корреляции Спирмена является непараметрическим методом определения тесноты связи между двумя рядами количественных показателей, приведенных к рангам.

Построение регрессионных моделей биовозраста с выделением предикторов. Для построения регрессионных моделей биовозраста использовались модели линейной регрессии, гребневой регрессии, решающие деревья, метод опорных векторов и композиции алгоритмов (случайный лес и бустинг). Выборка разбивалась на тестовую (20%) и обучающую (80%) части. Качество моделей оценивалось по коэффициенту детерминации. Адекватность моделей оценивалась по критерию Фишера [20].

Линейная регрессия. Модель линейной регрессии представляет собой взвешенную сумму факторов модели. В ходе построения модели необходимо вычислить веса каждого фактора в модели. При использовании методов машинного обучения веса модели вычисляются с помощью различных методов градиентного спуска на обучающей части выборки. В данной работе строились 2 линейные модели. Одна обучалась методом наименьших квадратов, вторая – методом стохастического градиентного спуска. Для оценки качества модели используется тестовая часть выборки, на которой с полученными значениями весов модели рассчитываются прогнозные значения целевой переменной. Рассчитанные значения сравниваются с истинными значениями целевой переменной, имеющимися в тестовой части и вычисляется функционал ошибки модели. В качестве функционала ошибки в модели регрессии чаще всего используют среднюю квадратичную ошибку.

Гребневая регрессия. Гребневая регрессия является разновидностью линейной регрессии, но в отличие от линейной регрессии при расчете весов модели в функционал ошибки вносится штрафное слагаемое для предотвращения неконтролируемого роста весов модели.

Решающие деревья. Деревья решений является непараметрическим методом обучения с учителем, используемым как для классификации, так и для регрессии. Они представляют собой иерархическую древовидную структуру, состоящую из решающих правил, вида «если, то». Структура содержит узлы и листья дерева. В узлах дерева помещаются решающие правила, а листьями дерева являются элементы выборки, для которых сработало решающее правило. Количество уровней иерархии полученной структуры дерева является глубиной дерева. При построении решающих деревьев можно строить дерево до тех пор, пока не останется элементов для разбиения или можно задать глубину до которой идет разбиение обучающего множества. Для задачи регрессии применяют алгоритм обучения CART (Classification and Regression Tree) [21].

Метод опорных векторов. Метод опорных векторов основан на том, что идет поиск поверхности, разделяющей выборку на классы (в случае классификации), или аппроксимирующей выборку (в случае регрессии). При применении метода опорных векторов для задач регрессии для нахождения весов модели используются функционал ошибки с меньшим, чем в гребневой регрессии штрафом за отклонения значения алгоритма от истинных значений [22]. Это достигается за счет использования кусочно-линейной функции ε-чувствительности в функционале ошибки вместо квадратичной.

Случайный лес. Это композиция алгоритмов, строящаяся на наборе решающих деревьев. Каждое из решающих деревьев это бинарное дерево, в узлах которого находится условие разделения выборки по заданному признаку. Каждое решающее дерево строится на своей подвыборке данных, созданной из основной выборки. Признаки на которых происходит разделение задаются случайным образом. Глубину построения каждого решающего дерева можно задать, либо оно строится до тех пор, пока в листе не останется один образец. Композиция алгоритмов случайного леса собирает ответы на всех решающих деревьях, входящих в него и результатом работы будет средний ответ по всем алгоритмам, входящих в композицию [23].

Градиентный бустинг. Эта композиция алгоритмов строится на любом наборе алгоритмов, но часто используются также решающие деревья. В отличии от случайного леса, каждый новый алгоритм, входящий в композицию обучается на том же наборе данных, но использует уже не ответы на выборке, а градиент ошибки предыдущего алгоритма. Результат работы композиции считается как взвешенная сумма ответов всех алгоритмов, входящих в композицию [23].

Выделение предикторов. Предикторами являются наиболее значимые показатели из анализируемого набора. Поскольку большинство примененных моделей были линейными и использовались масштабированные показатели, то вес перед признаком показывал значимость признака. На первом этапе был таким образом выделен базовый набор признаков. Затем отбирался такой набор признаков, который дает наилучшее качество модели [24].

Настройка модели. Настройка модели проводилась двумя способами. В первом способе велся поиск оптимального набора факторов, дающих максимальную точность модели. Для этого использовался метод Add-dell.

ADD-DELL метод. Метод предполагает пошаговое добавление признаков в модель и оценку ее качества на каждом шаге. Если добавляемый признак повышает качество модели, то он оставляется в наборе, если понижает, то удаляется из набора. В конечном итоге в наборе остаются только те признаки, которые дают наибольший вклад в качество модели [23].

Во втором способе для улучшения качества модели велась работа с данными. Первоначальная модель биологического возраста должна показывать биовозраст как можно ближе к реальному календарному возрасту для тех пациентов, у которых нет отклонений в функциональных данных. Поэтому для настройки первичной модели необходимо выделить из данных аномальные значения и обучать модель без них. Для выявления аномальных данных использовались следующие методы: метод опорных векторов с бинарной классификацией, метод изолирующего леса и метод ближайших соседей.

Метод изолирующего леса. Является разновидностью композиции алгоритмов Случайный лес и основан на идее, что выбросы будут попадать в листья дерева на первых этапах разделения решающего дерева. Таким образом те объекты, который были выделены в отдельные листья на первых шагах разделения решающего дерева и есть выбросы [25].

Метод ближайших соседей. Метод является инструментом кластеризации и позволяет выделить группы в выборке данных по принципу минимального расстояния между данными. Объект относится к тому кластеру, к которому относятся k его соседей. Количество соседей k задается в методе. Те объекты, которые невозможно отнести ни к одной группе, являются выбросами [21].

Инструменты. Все этапы работы с моделями (подготовка данных, удаление коллинеарных признаков, построение моделей, поиск предикторов) проводилась на платформе Anaconda с дистрибутивом Python 3.6 [26] с использованием различных библиотек. Для выявления коллинеарных признаков и оценки статистической значимости их корреляции использовалась библиотека SciPy [27]. Для построения моделей регрессии использовалась библиотека Scikit-learn [28]. В ней для работы с линейными моделями есть классы LinearRegression и SGDRegressor. Эти классы отличаются методами расчета весов модели, в первом используется метод наименьших квадратов, во втором – метод стохастического градиентного спуска. Оба класса имеют методы как для вычисления весов модели (метод fit()), так и для прогнозирования значений модели (метод predict()). Для реализации стохастического бустинга

использовалась библиотека XGBoost с API для языка Python 3.6 [29 https://xgboost.readthedocs.io/en/latest/index.html]. Для визуализации и обработки данных использовались библиотеки NumPy [30], pandas [31], matplotlib [32]. Для выявления аномалий использовались методы из библиотеки Scikit-learn – OneClassSVM для метода опорных векторов, IsolationForest для метода изолирующего леса и LocalOutlierFactor для метода ближайших соседей.

Результаты

Выявление коллинеарных признаков. На первом этапе исследований выделялись признаки, имеющие корреляцию между собой выше, чем с целевой переменной. Целевой переменной в исследовании служил календарный возраст пациента. Корреляция признаков между собой и с целевой переменной оценивалась с помощью коэффициента корреляции Пирсона. Результаты оценки корреляции признаков с целевой переменной приведены в Таблице 1.

Таблица 1 – оценка силы линейной взаимосвязи признаков с целевой переменной
Table 1 – the value of the power of the features linear correlation with target

Признак	Коэффициент корреляции Пирсона
АДС	0,27
АДД	0,05
АДП	0,32
ЗДВдох	-0,39
ЗДВыдох	-0,13
ЖЕЛ	-0,48
Масса тела	0,003
Аккомодация	-0,60
Острота слуха	0,60
Статическая балансировка	-0,53

Оценка статистической значимости корреляций признаков с целевой переменной с помощью критерия Стьюдента показала, что все признаки имеют статистически значимую корреляцию с целевой переменной. Как видно из Таблицы 1, наибольшую корреляцию с целевой переменной имеют три показателя: аккомодация, острота слуха и статическая балансировка.

Поскольку среди признаков есть группа показателей артериального давления (АДД, АДС, АДП) и группа показателей работы лёгких (ЗДВдох, ЗДВыдох, ЖЕЛ), то стоит ожидать, что показатели внутри каждой группы коррелируют между собой лучше, чем с целевой переменной. Результаты оценки корреляции показателей между собой приведены в Таблице 2.

Таблица 2 – матрица корреляций признаков
Table 2 – the matrix of the features correlation

	АДС	АДД	АДП	ЗДВдох	ЗДВыдох	ЖЕЛ
АДС	1,00	0,68	0,73	-0,07	-0,03	-0,07
АДД	0,68	1,00	0,05	-0,01	-0,01	0,02
АДП	0,73	0,05	1,00	-0,09	0,06	-0,10
ЗДВдох	-0,07	-0,01	-0,09	1,00	0,56	0,57
ЗДВыдох	-0,03	-0,01	0,06	0,56	1,00	0,25
ЖЕЛ	-0,07	0,02	-0,10	0,57	0,25	1,00

Как видно из Таблицы 1 и Таблицы 2, сильную корреляцию между собой, превышающую корреляцию с целевой переменной, имеют следующие пары признаков: АДС и АДД, АДС и АДП, ЗДВдох и ЗДВыдох, ЗДВдох и ЖЕЛ, ЗДВыдох и ЖЕЛ.

Построение регрессионных моделей биовозраста с выделением предикторов. На первом этапе для оценки весов факторов были построены модели для всех 10 параметров. Параметры в модель подавались после их масштабирования функцией scale из библиотеки Scikit-learn. Модели строились с помощью различных классов библиотеки Scikit-learn. Для построения линейной модели регрессии использовался класс LinearRegression() и SGDRegressor(), для гребенной регрессии – Ridge(), для ее модификации – BayesianRidge(), для метода опорных векторов – SVR(). Оценка точности модели проводилась с помощью средней абсолютной ошибки. Результаты приведены в Таблице 3.

Таблица 3 – средняя абсолютная ошибка (cao) и коэффициенты весов регрессионных моделей
Table 3 – the mae and weight coefficients of the regression models

	cao	Веса факторов									
		АДС	АДД	АДП	ЗДВдох	ЗДВыдох	ЖЕЛ	Масса тела	Акком.	ОС	СБ
Linear Regression	7,98	-1,07	1,10	1,94	-1,45	1,32	-3,46	0,15	-4,35	5,57	-2,60
SGD Regressor	7,99	-0,13	0,45	1,25	-1,43	1,36	-3,43	0,14	-4,34	5,52	-2,56
Ridge	7,98	-0,91	0,99	1,83	-1,45	1,31	-3,45	0,14	-4,34	5,55	-2,60
BayesianRidge	7,99	-0,23	0,54	1,38	-1,44	1,25	-3,39	0,12	-4,30	5,45	-2,60
SVR	8,08	-0,37	1,23	1,19	-1,68	1,23	-2,87	-0,42	-5,00	5,60	-1,91

Настройка модели.

Поиск оптимального набора признаков. Как видно из данных Таблицы 3, все модели выделяют четыре фактора – жизненная емкость легких, аккомодация, острота слуха и статическая балансировка. Эти факторы включены в базовый набор параметров для дальнейшего поиска оптимального набора параметров. Для поиска оптимального набора признаков использовался метод Add-dell на базовом наборе признаков. Результаты его работы показали (см. Таблицу 4), что оптимальным набором с использованием практически всех линейных моделей является следующий набор признаков: жизненная емкость легких, аккомодация, острота слуха, статическая балансировка, АДД, АДС и АДП (набор 1). При использовании SGDRegressor к набору добавляется еще масса тела (набор 2). При использовании композиций алгоритмов (Random Forest, XGBoosting) добавление параметра масса тела также улучшает качество модели.

Как видно по данным Таблицы 4 точность моделей на обоих наборах признаков не имеет сильного отличия, при этом модели не переобучаются (точность на обучающей части выборки не сильно отличается от точности на тестовой части). Использование композиций алгоритмов имеет тенденцию к переобучению моделей, но при этом дает более высокую точность получаемых моделей, чем применение линейных моделей.

Таблица 4 – точность моделей после отбора параметров
Table 4 – the accuracy of the regression models after parameters choose

Модель	Точность на наборе 1			Точность на наборе 2		
	R^2 на тестовой части	R2 на обучающей части	Средняя абсолютная ошибка	R2 на тестовой части	R2 на обучающей части	Средняя абсолютная ошибка
Linear Regressor	0,618	0,661	7,890	0,618	0,661	7,892
SGDRegressor	0,618	0,660	7,884	0,618	0,660	7,886
Ridge	0,618	0,661	7,887	0,618	0,661	7,889
BayesianRigde	0,618	0,661	7,880	0,618	0,661	7,882
SVR	0,615	0,656	7,912	0,610	0,655	7,957
Random Forest	0,681	0,942	6,890	0,695	0,943	6,890
XGBoosting	0,689	0,893	6,991	0,706	0,915	6,890

Поиск и удаление аномалий. Для поиска аномальных значений в выборке использовались следующие алгоритмы: метод опорных векторов, метод изолирующего леса и метод ближайших соседей.

Результаты работы регрессионных моделей с полным набором параметров на очищенных от аномалий выборках приведены в Таблице 5, с оптимальным набором параметров без массы тела (набор 1) в Таблице 6 и с массой тела (набор 2) в Таблице 7.

Таблица 5 – точность моделей после удаления аномалий с полным набором параметров
Table 5 – the accuracy of the regression models after anomalies deletion

Модель	Точность моделей после удаления аномалий								
	методом опорных векторов			методом изолирующего леса			методом ближайших соседей		
	R^2 на тестовой части	R2 на обучающей	Средняя абсолютная ошибка	R2 на тестовой части	R2 на обучающей	Средняя абсолютная ошибка	R2 на тестовой части	R2 на обучающей	Средняя абсолютная
Linear Regressor	0,610	0,661	7,700	0,590	0,623	6,917	0,643	0,650	7,741
SGDRegressor	0,609	0,661	7,723	0,591	0,626	6,921	0,644	0,650	7,738
Ridge	0,610	0,661	7,699	0,591	0,626	6,920	0,643	0,650	7,744
BayesianRigde	0,611	0,661	7,694	0,591	0,626	6,936	0,642	0,650	7,757
SVR	0,612	0,657	7,636	0,580	0,620	6,966	0,642	0,644	7,620
Random Forest	0,745	0,944	5,870	0,661	0,934	6,011	0,694	0,945	6,558
XGBoosting	0,721	0,906	6,241	0,634	0,990	6,238	0,702	0,940	6,686

Таблица 6 – точность моделей после удаления аномалий с набором параметров номер 1
Table 6 – the accuracy of the regression models after anomalies deletion

Модель	Точность моделей после удаления аномалий								
	методом опорных векторов			методом изолирующего леса			методом ближайших соседей		
	R^2 на тестовой части	R2 на обучающей части	Средняя абсолютная ошибка	R2 на тестовой части	R2 на обучающей части	Средняя абсолютная ошибка	R2 на тестовой	R2 на обучающей	Средняя абсолютная ошибка
Linear Regressor	0,610	0,661	7,700	0,582	0,624	7,055	0,652	0,644	7,651
SGD Regressor	0,609	0,661	7,711	0,583	0,624	7,053	0,652	0,643	7,671
Ridge	0,610	0,661	7,700	0,582	0,624	7,057	0,652	0,644	7,654
Bayesian Rigde	0,611	0,661	7,694	0,582	0,624	7,066	0,652	0,644	7,661
SVR	0,612	0,657	7,636	0,578	0,619	7,075	0,654	0,640	7,580
Random Forest	0,728	0,938	6,026	0,648	0,926	6,265	0,691	0,938	6,567
XGBoosting	0,721	0,906	6,241	0,660	0,927	6,353	0,711	0,907	6,572

Таблица 7 – точность моделей после удаления аномалий с набором параметров 2
Table 7 – the accuracy of the regression models after anomalies deletion

Модель	Точность моделей после удаления аномалий								
	методом опорных векторов			методом изолирующего леса			методом ближайших соседей		
	R^2 на тестовой части	R2 на обучающей части	Средняя абсолютная ошибка	R2 на тестовой части	R2 на обучающей части	Средняя абсолютная	R2 на тестовой части	R2 на обучающей	Средняя абсолютная ошибка
Linear Regressor	0,608	0,661	7,707	0,580	0,624	7,055	0,652	0,644	7,656
SGD Regressor	0,608	0,661	7,719	0,580	0,624	7,071	0,651	0,643	7,672
Ridge	0,608	0,661	7,706	0,580	0,624	7,057	0,652	0,644	7,658
Bayesian Rigde	0,610	0,661	7,700	0,580	0,624	7,066	0,652	0,644	7,665
SVR	0,610	0,657	7,641	0,572	0,621	7,116	0,653	0,638	7,575
Random Forest	0,750	0,934	5,917	0,643	0,932	6,249	0,701	0,943	6,412
XGBoosting	0,711	0,925	6,485	0,617	0,928	6,657	0,693	0,919	6,743

Как видно по данным Таблиц 5-7, наибольшую среднюю точность для всех моделей удается получить после удаления аномалий методом изолирующего леса. После удаления аномалий объем выборки сократился до 536 записей. В то же время применение

композиций алгоритмов после удаления аномалий методом опорных векторов дает наибольшую точность на всех наборах данных.

Обсуждение

Как видно по данным Таблиц 3 и 4, точность моделей после удаления коллинеарных признаков немного увеличивается, но дальнейшие исследования (см. Таблицы 5 - 7) показывают, что удаление аномалий из данных имеет большее влияние на повышение качества моделей, чем удаление коллинеарных признаков.

Максимальную точность показывает модель композиции случайного леса на полном наборе признаков после удаления аномалий методом опорных векторов. Но при этом наблюдается тенденция к переобучению модели. Минимальная средняя абсолютная ошибка модели, достигаемая исследованными методами машинного обучения, составила 5,87 лет.

Заключение

В работе было проведено исследование 7 методов построения регрессионных моделей (линейная регрессия, метод стохастического градиентного спуска, гребенная регрессия, Байесовская гребенная регрессия, метод опорных векторов, композиция алгоритмов случайный лес и композиция алгоритмов стохастический бустинг). Исследования проводились на трех наборах признаков – полный набор признаков, набор, отобранный методом ADD-Del и набор после Add-dell с добавленным признака «Масса тела». Кроме того, проводился поиск и удаление аномалий из данных. Применялось три способа поиска аномалий в данных (метод опорных векторов, метод изолирующего леса и метод ближайших соседей). Наибольшую точность показали модели, полученные на композиции алгоритмов, примененные к полному набору данных после удаления аномалий в данных методом опорных векторов.

Применение методов поиска аномалий в данных, используемых ранее в сфере машинного обучения, показало хорошие результаты и может быть успешно использовано в задачах построения медицинских прогностических моделей.

Благодарности

Работа выполнена в рамках государственного задания ФГБОУ ВО УГМУ Минздрава России на 2021 г. № 056-00054-21-00 от 17.12.2020 г., тема: «Индивидуализация подбора комплексной геропрофилактической терапии»" номер НИР 121030900298-9

Acknowledgments

The reported study was funded by state assignment of the Federal State Budgetary Educational Institution of Higher Education USMU of the Ministry of Health of Russia for 2021 No. 056-00054-21-00 dated December 17, 2020, topic: "Individualization of the selection of complex geroprophylactic therapy" the number 121030900298-9

ЛИТЕРАТУРА

1. L´opez-Ot´ın C., Blasco M.A., Partridge L., Serrano M., Kroemer G. The hallmarks of aging. *Cell* 2013;153(8):1194–1217. DOI: 10.1016/j.cell.2013.05.039
2. Kennedy B.K., Berger S.L., Brunet A., Campisi J., Cuervo A.M., Epel E.S., Franceschi C., Lithgow G.J., Morimoto R.I., Pessin J.E., Rando T.A., Richardson A., Schadt E.E., Wyss-Coray T., Sierra F. Geroscience: Linking Aging to Chronic Disease. *Cell.* 2014;159(4):709–713. DOI: 10.1016/j.cell.2014.10.039

3. Kirkwood T.B.L. Understanding the odd science of aging. *Cell.* 2005;120:437 – 447. DOI: 10.1016/j.cell.2005.01.027
4. Kirkwood T.B.L. Deciphering death: a commentary on Gompertz (1825) 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies'. *Philosophical Transactions Of The Royal Society Of London Series B.* 2015;370(1666):20140379–2014037. DOI: 10.1098/rstb.2014.0379
5. Yashin A.I., Arbeev K.G., Akushevich I., Kulminski A., Akushevich L., Ukraintseva S.V. Stochastic model for analysis of longitudinal data on aging and mortality. *Mathematical Biosciences.* 2007;208:538–551. DOI: 10.1016/j.mbs.2006.11.006
6. Yashin A.I., Arbeev K.G., Akushevich I., Kulminski A., Ukraintseva S.V., Stallard E., Land K.C. The quadratic hazard model for analyzing longitudinal data on aging, health, and the life span. *Physics of Life Reviews.* 2012;9:177–188. DOI: 10.1016/j.plrev.2012.05.002
7. Taneja S., Mitnitski A.B., Rockwood K., Rutenberg A.D. Dynamical network model for age-related health deficits and mortality. *Physical Review E* 2016;93(2):022309–022311. DOI: 10.1103/PhysRevE.93.022309
8. Farrell S.G., Mitnitski A.B., Rockwood K., Rutenberg A.D. Network model of human aging: Frailty limits and information measures. *Physical Review E* 2016;94(5):052409-052419. DOI: 10.1103/PhysRevE.94.052409
9. Farrell S, Mitnitski A, Rockwood K, Rutenberg A. Generating synthetic aging trajectories with a weighted network model using cross-sectional data. *Scientific Reports.* 2020;10(1):19833-19844. DOI: 10.1038/s41598-020-76827-3
10. Pierson E., Koh P.W., Hashimoto T., Koller D., Liang P. Inferring multidimensional rates of aging from cross-sectional data. Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019;89:97–107.
11. Putin E., Mamoshina P., Aliper A., Korzinkin M., Moskalev A., Kolosov A., Ostrovskiy A., Cantor C. Vijg J., Zhavoronkov A. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging (Albany NY).* 2016;8(5):1021-1033. DOI: 10.18632/aging.100968
12. Zhavoronkov A., Mamoshina P. Deep Aging Clocks: The Emergence of AI-Based Biomarkers of Aging and Longevity. *Trends Pharmacol Sci*. 2019;40(8):546-549. DOI: 10.1016/j.tips.2019.05.004
13. Levine ME. Assessment of Epigenetic Clocks as Biomarkers of Aging in Basic and Population Research. *J Gerontol A Biol Sci Med Sci.* 2020;75(3):463-465. DOI: 10.1093/gerona/glaa021.
14. Pyrkov T.V., Getmantsev E., Zhurov B., Avchaciov K., Pyatnitskiy M., Men'shikov, L., Khodova K., Gudkov A., Fedichev P. Quantitative characterization of biological age and frailty based on locomotor activity records. *Aging (Albany NY).* 2019;10:2973 - 2990. DOI: 10.1038/s41598-018-23534-9
15. Schultz M.B., Kane A.E., Mitchell S.J., MacArthur M.R., Warner E., Vogel D.S., Mitchell J.R., Howlett S.E., Bonkowski M.S., Sinclair D.A. Age and life expectancy clocks based on machine learning analysis of mouse frailty. *Nature Communications.* 2020;11(1):4618-4628. DOI: 10.1038/s41467-020-18446-0
16. Farrell S., Stubbings G., Rockwood K., Mitnitski A., Rutenberg A. The potential for complex computational models of aging. *Mechanisms of Ageing and Development*. 2020;193:111403-111418. DOI: 10.1016/j.mad.2020.111403
17. Zhavoronkov A., Mamoshina P., Vanhaelen Q., Scheibye-Knudsene M., Moskalev A., Alipera A. Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Research Reviews.* 2019;49:49-66. DOI: 10.1016/j.arr.2018.11.003

18. Fedintsev A., Daria Kashtanova D., Tkacheva O., Strazhesko I., Kudryavtseva A., Baranova A., Moskalev A. Markers of arterial health could serve as accurate non-invasive predictors of human biological and chronological age. *Aging.* 2017;9:1-13. DOI: 10.18632/aging.101227
19. Cohen A.A., Morissette-Thomas V., Ferrucci L., Fried L.P. Deep biomarkers of aging are population-dependent. *Aging (Albany NY)*. 2016;8(9):2253-2255. DOI: 10.18632/aging
20. Громыко Г.Л. *Теория статистики*. М.:ИНФРА-М, 2002
21. Aggarwal C.C. *Data Mining: The Textbook*. New York: Springer, 2015
22. Воронцов К. В. *Лекции по методу опорных векторов*. Доступно по: http://www.ccas.ru/voron/download/SVM.pdf (дата обращения 12.03.2021)
23. Лимановская О. В., Алферьева Т. И. *Основы машинного обучения: учебное пособие*. Екатеринбург: Издательство Уральского университета, 2020
24. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003;3: 1157–1182.
25. Liu F. T., Ting K. M., Zhou Z. Isolation Forest. *Eighth IEEE International Conference on Data Mining, Pisa, Italy*, 2008; 413-422. DOI: 10.1109/ICDM.2008.17
26. Платформа для обработки данных и машинного обучения Anaconda. Доступно по: https://www.anaconda.com (дата обращения 18.02.2021)
27. Библиотека SciPy. Достапно по: https://www.scipy.org/index.html (дата обращения 18.02.2021)
28. Faris H., Mafarja M.M., Heidari A.A., Aljarah I., Al-Zoubi A.M., Mirjalili S., Fujita H. An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems. *Knowledge-Based Systems*. 2018;154:43–67. DOI: 10.1016/j.knosys.2018.05.009
29. Библиотека XGBoost. Доступно по: https://xgboost.ai/ (дата обращения 17.02.2021)
30. Библиотека NumPy. Доступно по: https://numpy.org/ (дата обращения 18.02.2021)
31. Библиотека pandas. Доступно по: https://pandas.pydata.org/ (дата обращения 18.02.2021)
32. Библиотека Matplotlib. Доступно по: https://matplotlib.org/index.html (дата обращения 18.02.2021)

REFERENCES

1. L´opez-Ot´ın C., Blasco M.A., Partridge L., Serrano M., Kroemer G. The hallmarks of aging. *Cell* 2013;153:1194–1217. DOI: 10.1016/j.cell.2013.05.039
2. Kennedy B.K., Berger S.L., Brunet A., Campisi J., Cuervo A.M., Epel E.S., Franceschi C., Lithgow G.J., Morimoto R.I., Pessin J.E., Rando T.A., Richardson A., Schadt E.E., Wyss-Coray T., Sierra F. Geroscience: Linking Aging to Chronic Disease. *Cell*. 2014;159(4):709–713. DOI: 10.1016/j.cell.2014.10.039
3. Kirkwood T.B.L. Understanding the odd science of aging. *Cell*. 2005;120:437 – 447. DOI: 10.1016/j.cell.2005.01.027
4. Kirkwood T.B.L. Deciphering death: a commentary on Gompertz (1825) ‘On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies’. *Philosophical Transactions Of The Royal Society Of London Series B*. 2015;370(1666):20140379–2014037. DOI: 10.1098/rstb.2014.0379
5. Yashin A.I., Arbeev K.G., Akushevich I., Kulminski A., Akushevich L., Ukraintseva S.V. Stochastic model for analysis of longitudinal data on aging and mortality. *Mathematical Biosciences*. 2007;208:538–551. DOI: 10.1016/j.mbs.2006.11.006

6. Yashin A.I., Arbeev K.G., Akushevich I., Kulminski A., Ukraintseva S.V., Stallard E., Land K.C. The quadratic hazard model for analyzing longitudinal data on aging, health, and the life span. *Physics of Life Reviews*. 2012;9:177–188. DOI: 10.1016/j.plrev.2012.05.002
7. Taneja S., Mitnitski A.B., Rockwood K., Rutenberg A.D. Dynamical network model for age-related health deficits and mortality. *Physical Review E* 2016;93(2):022309–022311. DOI: 10.1103/PhysRevE.93.022309
8. Farrell S.G., Mitnitski A.B., Rockwood K., Rutenberg A.D. Network model of human aging: Frailty limits and information measures. *Physical Review E* 2016;94(5):052409-052419. DOI: 10.1103/PhysRevE.94.052409
9. Farrell S, Mitnitski A, Rockwood K, Rutenberg A. Generating synthetic aging trajectories with a weighted network model using cross-sectional data. *Scientific Reports*. 2020;10(1):19833-19844. DOI: 10.1038/s41598-020-76827-3
10. Pierson E., Koh P.W., Hashimoto T., Koller D., Liang P. Inferring multidimensional rates of aging from cross-sectional data. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* 2019;89:97–107.
11. Putin E., Mamoshina P., Aliper A., Korzinkin M., Moskalev A., Kolosov A., Ostrovskiy A., Cantor C. Vijg J., Zhavoronkov A. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging (Albany NY).* 2016;8(5):1021-1033. DOI: 10.18632/aging.100968
12. Zhavoronkov A., Mamoshina P. Deep Aging Clocks: The Emergence of AI-Based Biomarkers of Aging and Longevity. *Trends Pharmacol Sci.* 2019;40(8):546-549. DOI: 10.1016/j.tips.2019.05.004
13. Levine ME. Assessment of Epigenetic Clocks as Biomarkers of Aging in Basic and Population Research. *J Gerontol A Biol Sci Med Sci.* 2020;75(3):463-465. DOI: 10.1093/gerona/glaa021.
14. Pyrkov T.V., Getmantsev E., Zhurov B., Avchaciov K., Pyatnitskiy M., Men'shikov, L., Khodova K., Gudkov A., Fedichev P. Quantitative characterization of biological age and frailty based on locomotor activity records. *Aging (Albany NY).* 2019;10:2973 - 2990. DOI: 10.1038/s41598-018-23534-9
15. Schultz M.B., Kane A.E., Mitchell S.J., MacArthur M.R., Warner E., Vogel D.S., Mitchell J.R., Howlett S.E., Bonkowski M.S., Sinclair D.A. Age and life expectancy clocks based on machine learning analysis of mouse frailty. *Nature Communications.* 2020;11(1):4618-4628. DOI: 10.1038/s41467-020-18446-0
16. Farrell S., Stubbings G., Rockwood K., Mitnitski A., Rutenberg A. The potential for complex computational models of aging. *Mechanisms of Ageing and Development.* 2020;193:111403-111418. DOI: 10.1016/j.mad.2020.111403
17. Zhavoronkov A., Mamoshina P., Vanhaelen Q., Scheibye-Knudsene M., Moskalev A., Alipera A. Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Research Reviews.* 2019;49:49-66. DOI: 10.1016/j.arr.2018.11.003
18. Fedintsev A., Daria Kashtanova D., Tkacheva O., Strazhesko I., Kudryavtseva A., Baranova A., Moskalev A. Markers of arterial health could serve as accurate non-invasive predictors of human biological and chronological age. *Aging.* 2017;9:1-13. DOI: 10.18632/aging.101227
19. Cohen A.A., Morissette-Thomas V., Ferrucci L., Fried L.P. Deep biomarkers of aging are population-dependent. *Aging (Albany NY).* 2016;8(9):2253-2255. DOI: 10.18632/aging
20. Gromyko G.L. *Teoriya statistiki.* M .: INFRA-M, 2002.
21. Aggarwal C.C. *Data Mining: The Textbook.* New York: Springer, 2015
22. Vorontsov K. V. *Lektsii po metodu opornykh vektorov*. Available at: http://www.ccas.ru/voron/download/SVM.pdf (accessed 12.03.2021) (In Russ)

23. Limanovskaya O.V., Alferieva T.I. *Osnovy mashinnogo obucheniya: uchebnoye posobiye.* Yekaterinburg: Izdatel'stvo Ural'skogo universiteta, 2020
24. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003;3: 1157–1182.
25. Liu F. T., Ting K. M., Zhou Z. Isolation Forest. *Eighth IEEE International Conference on Data Mining*, 2008; 413-422. DOI: 10.1109/ICDM.2008.17
26. Anaconda - solutions for Data Science Practitioners and Enterprise Machine Learning. Available at: https://www.anaconda.com (accessed 18.02.2021)
27. SciPy library. Available at: https://www.scipy.org/index.html (accessed 18.02.2021)
28. Faris H., Mafarja M.M., Heidari A.A., Aljarah I., Al-Zoubi A.M., Mirjalili S., Fujita H. An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems. *Knowledge-Based Systems.* 2018;154;43–67. DOI: 10.1016/j.knosys.2018.05.009
29. XGBoost library. Available at: https://xgboost.ai/ (accessed 17.02.2021)
30. NumPy library. Available at: https://numpy.org/ (accessed 18.02.2021)
31. Pandas library. Available at: https://pandas.pydata.org/ (accessed 18.02.2021)
32. Matplotlib library. Available at: https://matplotlib.org/index.html (accessed 18.02.2021)

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Лимановская Оксана Викторовна кандидат химических наук, доцент кафедры интеллектуальных информационных технологий института фундаментального образования ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина», Екатеринбург, Российская Федерация
e-mail: *o.v.limanovskaia@urfu.ru*
ORCID: 0000-0002-2084-3916

Oksana Viktorovna Limanovskaya, Candidate of Chemical Sciences, Associate Professor of The Department of Intellectual Information Technologies, Institute of Fundamental Education, FSAEI HE «UrFU named after the first President of Russia B.N.Yeltsin», Ekaterinburg, Russian Federation

Гаврилов Илья Валерьевич, кандидат медицинских наук, доцент кафедры биохимии ФГБОУ ВО «Уральский государственный медицинский университет Минздрава РФ», Екатеринбург, Российская Федерация
e-mail: *iliagavrilov18@yandex.ru*
ORCID: 0000-0003-0806-1177

Iliya Valeriyavich Gavrilov, Candidate of Medical Sciences, Associate Professor of The Department of Biochemistry, FSBEI HE «USMU of the Ministry of Health of the Russian Federation», Ekaterinburg, Russian Federation

Мещанинов Виктор Николаевич, доктор медицинских наук, профессор, заведующий кафедрой биохимии ФГБОУ ВО «Уральский государственный медицинский университет Минздрава РФ», Екатеринбург, Российская Федерация
e-mail: *mv-02@yandex.ru*
ORCID: 0000-0001-7928-2503

Viktor Nikolaevich Meshchaninov, MD, Professor, Head of the Department of Biochemistry, Ural State Medical University of the Ministry of Health of the Russian Federation, Yekaterinburg, Russian Federation

Щербаков Денис Леонидович, кандидат биологических наук, ассистент кафедры биохимии ФГБОУ ВО «Уральский

Denis Leonidovich Shcherbakov, Candidate of Biological Sciences, Assistant of The Department of Biochemistry, FSBEI HE

государственный медицинский университет Минздрава РФ», Екатеринбург, Российская Федерация
e-mail: cdcom2@yandex.ru

«USMU of the Ministry of Health of the Russian Federation», Ekaterinburg, Russian Federation

Колос Елена Николаевна, магистрант кафедры интеллектуальных информационных технологий института фундаментального образования ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина», г. Екатеринбург, Российская Федерация
e-mail: iamhappluck@gmail.com

Elena Nikolaevna Kolos, Master's Student of the Department of Intellectual Information Technologies of the Institute of Fundamental Education of the First President of Russia B. N. Yeltsin Ural Federal University, Yekaterinburg, Russian Federation