УДК 51-76
DOI: 10.26102/2310-6018/2024.45.2.012

Построение гендерно- и возрастзависимых моделей оценки биовозраста на основе функциональных данных организма пациента

О.В. Лимановская[1], И.В. Гаврилов[1,2], В.Н. Мещанинов[1,2], А.С. Лисовенко[3]

[1]Институт медицинских клеточных технологий, Екатеринбург, Российская Федерация
[2]Уральский государственный медицинский университет Министерства здравоохранения Российской Федерации, Екатеринбург, Российская Федерация
[3]Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Екатеринбург, Российская Федерация

Резюме. Методы машинного обучения широко используются для построения медицинских прогностических моделей. В то же время, наряду с методами, основанными на классической статистике, применяются байесовские методы, которые наиболее эффективны при малых объемах выборки. В данной работе построен ряд моделей прогнозирования биовозраста пациента на основе его функциональных данных с использованием как классических методов машинного обучения, так и байесовского подхода. В качестве данных использовались результаты кластеризации, проведенной нами ранее в предыдущем исследовании на материале медицинских организаций "Свердловский областной клинический психоневрологический госпиталь для ветеранов войн" и «Институт медицинских клеточных технологий» за 1995–2022 гг. в объеме 6440 записи, где было получено 4 кластера, разделенных по полу и статусу пациента (стационарный и амбулаторный). Исходя из предположения, что пациенты в амбулаторном статусе имеют наименьшую разницу биологического и календарного возраста и, поэтому, вносят меньшую ошибку в точность модели, чем пациенты в стационарном статусе, принято решение строить модели только для пациентов в амбулаторном статусе. В работе построен набор моделей для 2 кластеров – кластера мужчин в амбулаторном статусе (объем выборки 344 записи) и кластера женщин в амбулаторном статусе (объем выборки 991 запись). Анализ распределения возраста в каждой группе показал двумодальное распределение с границей при значении 40 лет. Поэтому группы были разделены по возрасту на две части: до 40 лет и после. Для выбора классических моделей машинного обучения использовалась платформа lazypredict. Для каждой группы выбирались 4 метода, дающие наибольшую точность и строились модели на их основе, а также использовались ансамбли моделей – stacking и votinmg. Точность моделей на тестовых данных составила от 4,1 до 6,3 лет. В байесовском подходе построена линейная многофакторная модель регрессии с заданным априорным распределением коэффициентов регрессии. Точность моделей составила от 4,9 до 6,6 лет.

Ключевые слова: байесовский подход, случайный лес, ансамбли моделей, voiting, stacking, геропрофилактическое воздействие, прогнозирование эффективности лечения, биовозраст.

Для цитирования: Лимановская О.В., Гаврилов И.В., Мещанинов В.Н., Лисовенко А.С. Построение гендерно- и возрастзависимых моделей оценки биовозраста на основе функциональных данных организма пациента. *Моделирование, оптимизация и информационные технологии.* 2024;12(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1583 DOI: 10.26102/2310-6018/2024.45.2.012

Building gender- and age-dependent models for assessing bio-age based on the functional data of the patient's body

O.V. Limanovskaya[1], I.V. Gavrilov[1,2], V.N. Meshchaninov[1,2], Lisovenko A.S.[3]

[1]*Institute of Medical Cell Technologies, Ekaterinburg, the Russian Federation*
[2]*Ural State Medical University of the Ministry of Health of the Russian Federation, Ekaterinburg, the Russian Federation*
[3]*Ural Federal University named after the first President of Russia B.N. Yeltsin, Ekaterinburg, the Russian Federation*

Abstract: Machine learning methods are widely used to build medical predictive models. At the same time, along with methods based on classical statistics, Bayesian methods are used, which are most effective for small sample sizes. In this paper, a number of models for predicting the patient's bio-age based on his functional data using both classical machine learning methods and the Bayesian approach are constructed. The data used were the results of clustering that we carried out earlier in a previous study on the material of medical organizations "Sverdlovsk Regional Clinical Psychoneurological Hospital for War Veterans" and "Institute of Medical Cell Technologies" for 1995–2022 in a volume of 6440 records, where 4 clusters were obtained, divided by gender and patient status (inpatient and outpatient). Based on the assumption that patients in outpatient status have the smallest difference in biological and calendar age, and therefore make less error in the accuracy of the model than patients in inpatient status, it was decided to build models only for patients in outpatient status. The work constructed a set of models for 2 clusters – a cluster of men in outpatient status (sample size 344 records) and a cluster of women in outpatient status (sample size 991 records). The analysis of the age distribution in each group showed a two-modal distribution with a boundary at a value of 40 years. Therefore, the groups were divided by age into two parts: up to 40 years and after. The lazypredict platform was used to select classical machine learning models. For each group, 4 methods were selected that gave the highest accuracy and models were built based on them, as well as ensembles of models - stacking and votinmg. The accuracy of the models based on the test data ranged from 4.1 to 6.3 years. In the Bayesian approach, a linear multifactorial regression model with a given a priori distribution of regression coefficients is constructed. The accuracy of the models ranged from 4.9 to 6.6 years.

Keywords: Bayesian approach, random forest, ensembles of models, voiting, stacking, geroprophylactic effect, predicting the effectiveness of treatment, bio-growth.

For citation: Limanovskaya O.V., Gavrilov I.V., Meshchaninov V.N., Lisovenko A.S. Building gender- and age-dependent models for assessing bio-age based on the functional data of the patient's body. *Modeling, Optimization and Information Technology*. 2024;12(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1583 DOI: 10.26102/2310-6018/2024.45.2.012 (In Russ.).

Введение

Методы машинного обучения и искусственного интеллекта нашли широкое применение в построении прогностических моделей в медицинской практике [1–6]. Как показано в обзорной работе [6], наиболее часто среди классических методов машинного обучения, применяемых для построения прогностических моделей, используются ансамбли решающих деревьев – случайный лес (Random Forest) и градиентный спуск (XGBoost). Тем не менее, при небольших объемах выборки, методы байесовской статистики имеют преимущества перед классическими алгоритмами машинного обучения и дают более высокую точность. Такое преимущество достигается за счет использования как априорной информации об исследуемом объекте, так и статистических данных по результатам исследований. Эта особенность байесовского подхода дала возможность его применения для построения медицинских диагностических моделей. Так в работе [7] построена система компьютерной диагностики вероятностными методами, применяемая для дифференциальной диагностики механической и паренхиматозной желтухи. В работе показан способ учета медицинских знаний в математическом подходе при построении диагностической системы. Экспертные знания о болезни позволяют получить априорные распределения

диагностических признаков, которые затем используются для расчета вероятности заболевания на основе байесовского подхода.

В данной работе для построения гендерно-зависимых прогностических моделей для определения биовозраста, наряду с классическими методами машинного обучения и ансамблей моделей, применен байесовский подход. Поскольку для построения гендерно-зависимых моделей биовозраста используется не вся база данных, а только ее части, то объем выборок заметно снижается, особенно в случае построения моделей для прогнозирования биовозраста мужчин. Объем выборки для них составил всего 344 записи. В связи с уменьшением объема выборки становится актуальным использование байесовского подхода.

Материалы и методы

Данные. Для построения моделей использованы результаты кластеризации пациентов, проведенной нами ранее в работе [8]. Как было показано в работе [8], разделение пациентов на кластеры произошло по полу и статусу пациента. Получено 4 кластера – 2 мужских кластера с амбулаторным и стационарным статусами пациентов и 2 женских кластера с амбулаторным и стационарным статусом пациентов. Исходя из предположения, что пациенты с амбулаторным статусом имеют меньшее отличия в календарном возрасте, и, следовательно, вносят меньшее влияние в ошибку модели, чем пациенты со стационарным статусом, принято решение использовать для построения моделей только пациентов с амбулаторным статусом. Таким образом для построения моделей использовались данные 2 кластеров – мужского кластера пациентов с амбулаторным статусом объемом 344 записи и женского кластера пациентов с амбулаторным кластером объемом 991 запись. Анализировались 13 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 – время задержки дыхания на вдохе в секундах.
5. ЗДВыдох – время задержки дыхания на выдохе в секундах.
6. ЖЕЛ – жизненная емкость легких в мл.
7. Масса – масса тела в кг.
8. Аккомодация – аккомодация хрусталика глаза в диоптриях.
9. Острота слуха – острота слуха в децибелах при частоте звуковых колебаний 4000 Гц.
10. Стат. балансировка – статическая балансировка в секундах.

Классические методы построения прогностических моделей. При построении прогностических моделей использовался набор следующих классических методов: случайный лес, градиентный бустинг и метод опорных векторов. Также применялся ансамбль из этих моделей (использовался Стекинг и Voiting).

Метод опорных векторов. В основе метода опорных векторов заложен поиск поверхности, разделяющей выборку на классы (в случае классификации), или аппроксимирующей выборку (в случае регрессии) [9].

Случайный лес. Этот метод представляет собой композицию решающих деревьев, представляющих собой бинарные деревья, в узлах которых находится условие разделения выборки по заданному признаку. Результатом работы метода является усредненный ответ по всем решающим деревьям, входящим в него [10].

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 [10].

Стекинг. Это ансамбль моделей, в который можно включить любое число любых моделей, при этом каждая модель в ансамбле делает свое предсказание на выборке и по ним заполняется мета-матрица, которая используется для обучения мета-алгоритма [11].

Voiting. Это также ансамбль моделей, но, в отличии от стекинга, итоговое предсказание считается как среднее по всем предсказаниям моделей, входящих в ансамбль.

Реализация алгоритмов выполнена на языке Python3.11 на фреймворке Anaconda с использованием библиотеки sklearn.

Результаты

На первом этапе исследования было оценено распределение параметра «возраст» в обеих гендерных группах данных. Результаты оценки представлены на Рисунках 1 и 2.

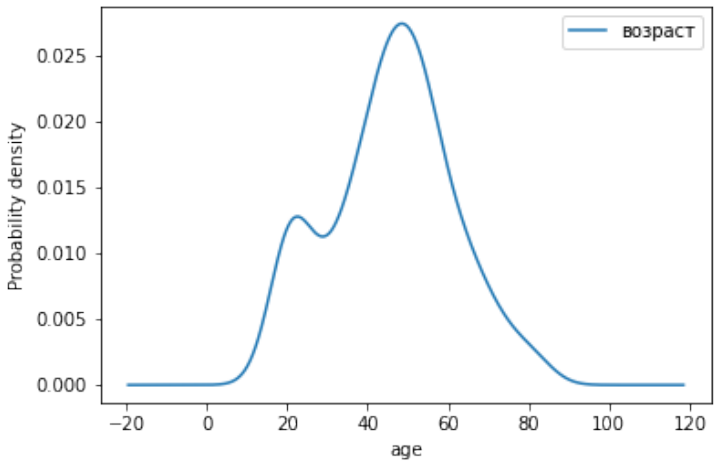

Рисунок 1 – Распределение по плотности вероятности параметра «возраст» в мужском кластере
Figure 1 – The value of age parameter distribution in man cluster

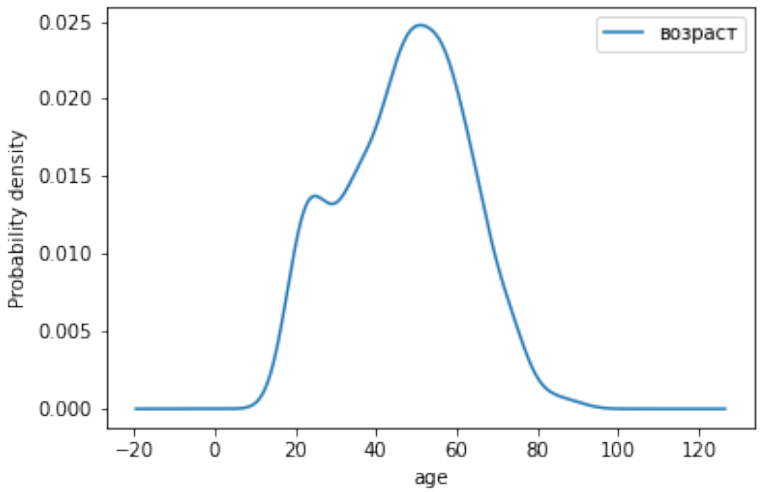

Рисунок 2 – Распределение по плотности вероятности параметра «возраст» в женском кластере
Figure 2 – The value of age parameter distribution in woman cluster

Как видно из Рисунков 1 и 2, распределение параметра «возраст» для обеих групп имеет двумодальный характер – группы можно разделить по возрасту на 2 части – моложе 40 лет и старше 40 лет, включая в старшую возрастную группу пациентов 40 лет. Таким образом, обе группы данных были разделены по возрасту на 2 подгруппы каждая и в результате все модели строились для 4 групп – мужской кластер возрастной категории до 40 лет (110 записей), мужской кластер возрастной категории старше 40 лет включительно (234 записи), женский кластер возрастной категории до 40 лет (308 записи), женский кластер возрастной категории старше 40 лет включительно (683 записи).

Использование байесовского подхода. Байесовский подход использовался для построения линейной модели регрессии. При построении модели линейной регрессии предполагалось, что коэффициенты линейной регрессии имеют распределение, схожее с распределением параметров, чьими коэффициентами они являются. Таким образом, на первом этапе построения байесовской модели линейной регрессии, нужно было получить информацию о априорном распределении параметров датасета. Для этих целей использовалась оценка параметров распределения.

Оценка распределений параметров в данных. Для оценки параметров распределения данных была написана функция dist_norm, код которой представлен ниже:

```
def dist_norm(mu, sigma, df):
    t_dist = pm.Normal.dist(mu = mu, sigma = sigma)
    x_eval = np.linspace(-6, 6, 300)
    plt.plot(x_eval, pm.math.exp(pm.logp(t_dist, x_eval)).eval(), label="Normal", lw=2.0)
    df.plot.kde()
    plt.xlabel("x")
    plt.ylabel("Probability density")
    plt.legend()
```

В этой функции с помощью метода kde() из библиотеки pandas строится распределение заданного параметра из данных. Далее на этом же графике методами библиотеки pymc и matplotlab строится нормальное распределение с заданными параметрами. Параметры нормального распределения задаются при вызове функции.

До построения распределения параметр нормализовался с помощью метода StandardScaler() библиотеки sklearn, после чего данные приводились к размерности от 0 до 10.

Эта функция была применена для каждого параметра во всех группах данных. Параметры распределения, а именно значения математического ожидания и дисперсии, подбирались вручную до получения визуально схожего графика нормального распределения, как показано на Рисунке 3. На Рисунке 3 в качестве примера показан результат оценки распределения параметра «Жизненная емкость лёгких» в женской группе до 40 лет.

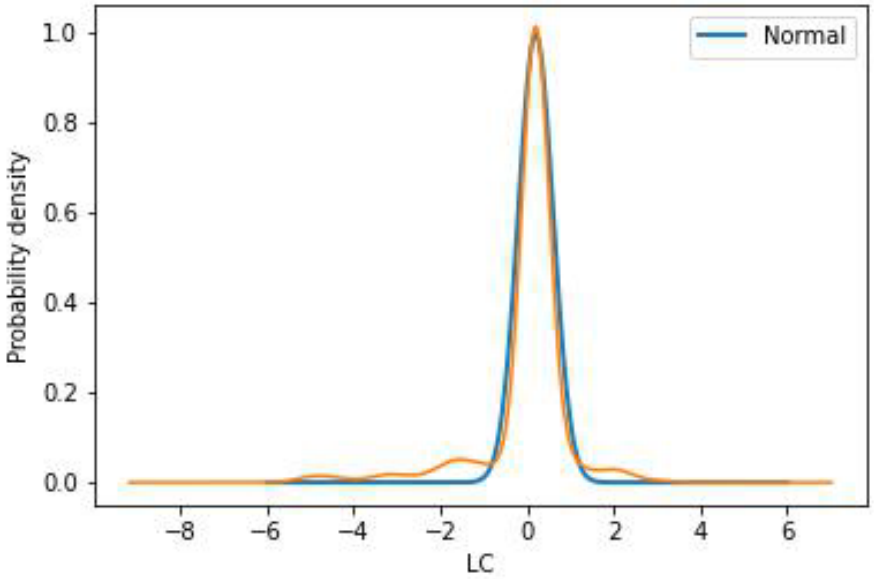

Рисунок 3 – Результат оценки распределения параметра «Жизненная емкость лёгких» в женской группе до 40 лет
Figure 3 – The result of distribution value of lang capacity parameter in women cluster under 40 years old

Построение линейной модели регрессии. Полученные на этапе оценки параметров распределения значения математических ожиданий и дисперсий использовались для построения априорных распределений коэффициентов линейной модели. Априорные распределения строились с помощью метода Normal библиотеки pymc.

На основе полученных априорных распределений коэффициентов линейной регрессии и данных датасетов рассчитывалось значение математического ожидания целевой переменной – возраста по формуле линейной регрессии (1).

$$\begin{gathered}\text{mu} = \text{epsilon} + \text{ads}\cdot\text{x\_train['АДС']} + \text{add}\cdot\text{x\_train['АДД']} + \text{adp}\cdot\text{x\_train['АДП']} + \\ \text{zdin}\cdot\text{x\_train['ЗДВдох']} + \text{zdout}\cdot\text{x\_train['ЗДВыдох']} + \text{gel}\cdot\text{x\_train['ЖЕЛ']} + \text{massa}\times \\ \times\text{x\_train['Масса тела']} + \text{akkom}\cdot\text{x\_train['Аккомодация']} + \text{sluh}\cdot\text{x\_train['Острота} \\ \text{слуха']} + \text{balance}\cdot\text{x\_train['Стат.балансировка']},\end{gathered} \quad (1)$$

где epsilon – свободный член линейной регрессии; ads, add, adp, zdin, zdout, gel, massa, akkom, sluh, balance – коэффициенты линейной регрессии при параметрах АДС, АДД, АДП, ЗДВдох, ЗДВыдох, ЖЕЛ, Масса тела, Аккомадация, Острота слуха, Стат. Балансировка, соответственно; x_train['АДС'], x_train['АДД'], x_train['АДП'], x_train['ЗДВдох'], x_train['ЗДВыдох'], x_train['ЖЕЛ'], x_train['Масса тела'], x_train['Аккомодация'], x_train['Острота слуха'], x_train['Стат.балансировка'] – векторы значений параметров из обучающей части выборки.

Далее строилось постриорное распределение целевой переменной с использованием рассчитанного значения математического ожидания и заданной дисперсии, равной 0,95. Из полученного распределения методом sample библиотеки pymc собиралась выборка из 2000 объектов. На основе этой выборки рассчитывались средние значения коэффициентов линейной регрессии

Используя полученные коэффициенты линейной регрессии и тестовые данные по формуле (2) рассчитывался прогноз биовозраста.

$$\begin{aligned} bioage = epsilon\_mean + ads\_mean \cdot x\_test['АДС'] + add\_mean \cdot x\_test['АДД'] + \\ adp\_mean \cdot x\_test['АДП'] + zdin\_mean \cdot x\_test['ЗДВдох'] + \\ zdout\_mean \cdot x\_test['ЗДВыдох'] + gel\_mean \cdot x\_test['ЖЕЛ'] + massa\_mean \times \\ \times x\_test['Масса\ тела'] + akkom\_mean \cdot x\_test['Аккомодация'] + \\ sluh\_mean \cdot x\_test['Острота\ слуха'] + balance\_mean \cdot x\_test['Стат.балансировка'], \end{aligned} \quad (2)$$

где epsilon_mean – среднее значение свободного члена линейной регрессии, определенное из выборки постприорного распределения целевой переменной; ads_mean, add_mean, adp_mean, zdin_mean, zdout_mean, gel_mean, massa_mean, akkom_mean, sluh_mean, balance_mean – средние значения коэффициентов линейной регрессии при параметрах АДС, АДД, АДП, ЗДВдох, ЗДВыдох, ЖЕЛ, Масса тела, Аккомадация, Острота слуха, Стат. Балансировка, соответственно, определенные из выборки постприорного распределения целевой переменной; x_test['АДС'], x_test['АДД'], x_test['АДП'], x_test['ЗДВдох'], x_test['ЗДВыдох'], x_test['ЖЕЛ'], x_test['Масса тела'], x_test['Аккомодация'], x_test['Острота слуха'], x_test['Стат.балансировка'] – векторы значений параметров из тестовой части выборки.

Для оценки качества модели использовался коэффициент детерминации r^2 и абсолютная квадратичная ошибка (mae). В тестовую часть выделялось 20% от общей выборки. Оценки линейных моделей для всех 4 групп приведены в Таблице 1.

Таблица 1 – Оценки линейных моделей
Table 1 – The values of linear models

Группа	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Женский кластер моложе 40 лет	0,21	4,87	0,11	5,15
Женский кластер старше 40 лет	0,24	6,58	0,25	6,84
Мужской кластер моложе 40 лет	0,03	5,83	0,22	5,40
Мужской кластер старше 40 лет	0,27	6,58	0,37	6,41

Классические модели машинного обучения. Помимо байесовского подхода для каждой из 4 групп строились классические модели машинного обучения. Выбор моделей был сделан на основе результатов быстрого сканирования моделей, проведенного с помощью проекта lazypredict. Для каждой группы проводилось свое сканирование моделей и выбирались топ-4 по качеству модели.

Модели, построенные для данных по мужской группе старше 40 лет.

В результате сканирования моделей во фреймворке lazypredict выделены 4 модели, имеющие наилучшее качество на тестовой выборке. В этой группе пациентов наилучшими моделями для прогнозирования биовозраста оказались модели случайного леса (RandomForestRegressor), сверхслучайных деревьев (ExtraTressRegressor), градиентного бустинга (GradientBoostRegressor) и модель гамма-регрессора (GammaRegressor).

Для каждой из этих моделей была проведена настройка параметров с целью получения наилучшего качества на тестовых данных. Для моделей, построенных на решающих деревьях (случайный лес, сверхслучайные деревья, градиентный бустинг),

настраивались глубина построения дерева и число деревьев. Результаты настройки приведены в Таблице 2. Оценки качества полученных моделей на тестовой выборке приведены в Таблице 3.

Таблица 2 – Настройка параметров моделей
Table 2 – Setting model parameters

Параметр настройки	Случайный лес	Сверхслучайные деревья	Градиентный бустинг
Глубина деревьев	5	5	2
Число деревьев	5	290	40

Таблица 3 – Оценки моделей машинного обучения для данных мужского кластера старше 40 лет
Table 3 – The values of machine learning models for data of man cluster up to 40 years old

Модель	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Случайный лес	0,42	5,80	0,63	4,49
Сверхслучайные деревья	0,44	5,52	0,57	5,21
Градиентный бустинг	0,37	5,87	0,56	5,34
Гамма-регрессор	0,32	6,34	0,38	6,31

Модели, построенные для данных по мужской группе моложе 40 лет. В этой группе в результате сканирования моделей во фреймворке lazypredict также выделены 4 модели, имеющие наилучшее качество на тестовой выборке: модель случайного леса, сверхслучайных деревьев, адабустинга (AdaBoostRegressor) и модель хист-градиентного регрессора (HistGradientRegressor).

Для моделей, построенных на решающих деревьях (случайный лес, сверхслучайные деревья, адабустинг), настраивалась глубина построения дерева и число деревьев. Результаты настройки приведены в Таблице 4. Оценки качества полученных моделей на тестовой выборке приведены в Таблице 5.

Таблица 4 – Параметры моделей на решающих деревьях, настроенных для группы данных мужчин моложе 40 лет
Table 4 – The parameters of models on decision trees configured for a group of men under 40 years of age

Параметр настройки	Случайный лес	Сверхслучайные деревья	Адабустинг
Глубина деревьев	2	5	5
Число деревьев	20	20	35

Таблица 5 – Оценки моделей машинного обучения для данных мужского кластера моложе 40 лет
Table 5 – The values of machine learning models for data of man cluster under 40 years old

Модель	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Случайный лес	0,07	5,61	0,51	4,49
Сверхслучайные деревья	0,12	5,42	0,67	3,00
Адабустинг	0,16	4,87	0,68	2,57
Хистбустингрегрессор	0,10	5,59	0,57	3,80

Модели, построенные для данных по женской группе старше 40 лет. В этой группе в результате сканирования моделей во фреймворке lazypredict выделены 4 модели, имеющие наилучшее качество на тестовой выборке: модель случайного леса, сверхслучайных деревьев, градиентного бустинга и модель хист-градиентного регрессора.

Для моделей, построенных на решающих деревьях (случайный лес, сверхслучайные деревья, градиентный бустинг), настраивалась глубина построения дерева и число деревьев. Результаты настройки приведены в Таблице 6. Оценки качества полученных моделей на тестовой выборке приведены в Таблице 7.

Таблица 6 – Параметры моделей на решающих деревьях, настроенных для группы данных женщин старше 40 лет
Table 6 – Parameters of decision trees models tuned for a data set of women over 40 years of age

Параметр настройки	Случайный лес	Сверхслучайные деревья	Градиентный бустинг
Глубина деревьев	10	20	4
Число деревьев	60	20	55

Таблица 7 – Оценки моделей машинного обучения для данных женского кластера старше 40 лет
Table 7 – Estimates of machine learning models for data from the woman cluster over 40 years of age

Модель	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Случайный лес	0,30	6,23	0,61	4,50
Сверхслучайные деревья	0,30	6,24	0,71	2,31
Градиентный бустинг	0,30	6,24	0,57	4,73
Хистбустингрегрессор	0,26	6,36	0,62	4,29

Модели, построенные для данных по женской группе моложе 40 лет. В этой группе в результате сканирования моделей во фреймворке lazypredict выделены 4 модели, имеющие наилучшее качество на тестовой выборке: модель случайного леса, градиентного бустинга и Ридж регрессии (Ridge).

Для моделей, построенных на решающих деревьях (случайный лес, градиентный бустинг), настраивалась глубина построения дерева и число деревьев. Результаты

настройки приведены в Таблице 8. Оценки качества полученных моделей на тестовой выборке приведены в Таблице 9.

Таблица 8 – Параметры моделей на решающих деревьях, настроенных для группы данных женщин моложе 40 лет
Table 8 – The parameters of models on decision trees configured for a group of women under 40 years of age

Параметр настройки	Случайный лес	Градиентный бустинг
Глубина деревьев	3	3
Число деревьев	225	40

Таблица 9 – Оценки моделей машинного обучения для данных женского кластера моложе 40 лет
Table 9 – The values of machine learning models for data of woman cluster under 40 years old

Модель	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Градиентный бустинг	0,23	4,83	0,38	4,11
Случайный лес	0,24	4,88	0,27	4,62
Ридж регрессор	0,19	4,94	0,11	5,16

Ансамбли классических моделей машинного обучения. Для данных каждой из 4 групп пациентов были построены ансамбли из ранее выбранных моделей машинного обучения. Для построения ансамблей использовался как метод голосования – voiting, так и stacking.

Ансамбли моделей, построенные для данных по мужской группе старше 40 лет. Для данной группы использовалось несколько вариантов ансамблей моделей. Во-первых, был апробирован ансамбль на основе voiting всех ранее выбранных моделей для этой группы – случайного леса, сверхслучайных деревьев, градиентного бустинга и модель гамма-регрессора. Все модели использовались с настройками, найденными ранее для данных этой группы пациентов, указанных в Таблице 2. Далее использовался stacking с разными наборами моделей, но в качестве мета-алгоритма во всех вариантах использовался гамма-регрессор. В первом варианте stacking применялся для набора из 2 моделей – случайного леса, сверхслучайных деревьев. Он обозначен в Таблице 10 как stacking1. Во втором варианте stacking применялся для набора из 3 моделей – случайного леса, сверхслучайных деревьев, градиентного бустинга. Он обозначен в Таблице 10 как stacking2. В третьем варианте в качестве набора моделей использовались все 4 модели – случайного леса, сверхслучайных деревьев, градиентного бустинга и модель гамма-регрессора. Оценки ансамблей моделей представлены в Таблице 10.

Таблица 10 – Оценки ансамблей моделей машинного обучения для данных мужского кластера старше 40 лет
Table 10 – Ensemble estimates of machine learning models for data from the male cluster over 40 years of age

Ансамбль моделей	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Voiting	0,46	5,56	0,58	5,20
Stacking1	0,45	5,48	0,59	5,06
Stacking2	0,48	5,35	0,62	4,78
Stacking3	0,40	5,89	0,45	5,89

Ансамбли моделей, построенные для данных по мужской группе моложе 40 лет.
Для данной группы также использовалось несколько вариантов ансамбли моделей: voiting на основе voiting всех ранее выбранных моделей для этой группы – случайного леса, сверхслучайных деревьев, адабустинга и модель хист-градиентного спуска и три варианта stacking с хист-градиентный спуском в качестве мета алгоритма. В первом варианте stacking применялся для набора из 2 моделей – случайного леса, сверхслучайных деревьев. Он обозначен в Таблице 11 как stacking1. Во втором варианте stacking применялся для набора из 3 моделей – случайного леса, сверхслучайных деревьев, адабустинга. Он обозначен в Таблице 11 как stacking2. В третьем варианте (stacking3) в качестве набора моделей использовались все 4 ранее описанные модели. Оценки ансамблей моделей представлены в Таблице 11. Все модели использовались с настройками, найденными ранее для данных этой группы пациентов, указанных в Таблице 4.

Таблица 11 – Оценки ансамблей моделей машинного обучения для данных мужского кластера старше 40 лет
Table 11 – Ensemble estimates of machine learning models for data from the male cluster over 40 years of age

Ансамбль моделей	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Voiting	0,15	5,32	0,64	3,35
Stacking1	-0,01	6,10	0,52	4,24
Stacking2	0,02	6,00	0,53	4,17
Stacking3	0,07	5,83	0,51	4,30

Ансамбли моделей, построенные для данных по женской группе старше 40 лет.
Для данной группы также использовалось несколько вариантов ансамблей моделей: voiting на основе всех ранее выбранных моделей для этой группы – случайного леса, сверхслучайных деревьев, градиентного бустинга, модель хист-градиентного спуска и три варианта stacking с хист-градиентным спуском в качестве мета-алгоритма. В первом варианте stacking применялся для набора из 2 моделей – случайного леса, сверхслучайных деревьев. Он обозначен в Таблице 12 как stacking1. Во втором варианте stacking применялся для набора из 3 моделей – случайного леса, сверхслучайных деревьев, градиентного бустинга. Он обозначен в Таблице 12 как stacking2. В третьем варианте (stacking3) в качестве набора моделей использовались все 4 ранее описанные модели. Оценки ансамблей моделей представлены в Таблице 12. Все модели

использовались с настройками, найденными ранее для данных этой группы пациентов, указанных в Таблице 6.

Таблица 12 – Оценки ансамблей моделей машинного обучения для данных женского кластера старше 40 лет
Table 12 – Estimating ensembles of machine learning models for women cluster data over 40 years of age

Ансамбль моделей	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Voiting	0,32	6,12	0,54	4,94
Stacking1	0,21	6,65	0,54	4,94
Stacking2	0,20	6,78	0,51	5,21
Stacking3	0,11	7,16	0,44	5,74

Ансамбли моделей, построенные для данных по женской группе моложе 40 лет.
Для данной группы также использовалось несколько вариантов ансамблей моделей: voiting на основе всех ранее выбранных моделей для этой группы – случайного леса, градиентного бустинга и Ридж регрессора и два варианта stacking с Ридж регрессором в качестве мета алгоритма. В первом варианте stacking применялся для набора из 2 моделей – случайного леса и градиентного бустинга. Он обозначен в таблице 13 как stacking1. Во втором варианте (stacking2) stacking применялся для набора из всех ранее описанных моделей. Оценки ансамблей моделей представлены в Таблице 13. Все модели использовались с настройками, найденными ранее для данных этой группы пациентов, указанных в Таблице 8.

Таблица 12 – Оценки ансамблей моделей машинного обучения для данных женского кластера моложе 40 лет
Table 12 – The values of ensembles of machine learning models for data of woman cluster under 40 years old

Ансамбль моделей	Оценки на тестовой части выборки		Оценки на обучающей части выборки	
	R^2	mae	R^2	mae
Voiting	0,23	4,86	0,27	4,61
Stacking1	0,24	4,89	0,22	4,80
Stacking2	0,24	4,89	0,22	4,81

Обсуждение

Результаты показали, что применение байесовского подхода для построения моделей прогнозирования биовозраста для всех групп дает модули с низкой точностью, не способные объяснить дисперсию данных (R^2 на тестовых данных меняется от -0,01 до 0,27). Поэтому для построения системы предсказания биовозраста решено использовать модели, построенные классическими методами машинного обучения. Так, в группе данных по мужчинам старше 40 лет наибольшую точность показали модель сверхслучайных деревьев и stacking3, включающие в себя в качестве моделей все 4 выбранных ранее модели. Но модель stacking3 давала большее переобучение, чем модель сверхслучайных деревьев, поэтому для построения системы прогнозирования биовозраста решено выбрать модель сверхслучайных деревьев. В группе данных мужчин моложе 40 лет наибольшую точность показала модель Ада бустинга. В группе данных

женщин старше 40 лет высокую точность показали модели случайного леса, сверхслучайных деревьев и градиентного бустинга (для всех моделей R^2 = 0,30), наибольшую точность дал ансамбль моделей voting R^2=0,32. Но при этом менее всех переобучилась модель градиентного бустинга, поэтому ее и решено взять для работы в системе. В группе данных по женщинам моложе 40 лет практически все классические модели показали примерно одинаковую точность – R^2 менялось от 0,23 до 0,24. Тем не менее, наименьшее переобучение показала модель случайного леса, которую и решено было взять в работу в системе предсказания биовозраста.

Заключение

В работе исследованы различные методы построения моделей прогнозирования биовозраста, использовались как классические методы машинного обучения и ансамбли из них, так и байесовский подход. Результаты показали, что байесовский подход не дает существенного повышения качества модели, что, возможно, связано с неточным знанием априорного распределения параметров модели. Наилучшие результаты в построении моделей прогнозирования биовозраста показали классические модели, которые дают хорошую точность в пределах от 4 до 6 лет и не переобучаются. Поэтому для построения системы прогнозирования биовозраста с целью выявления механизма старения будут использованы классические модели машинного обучения.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Sidey-Gibbons J.A.M., Sidey-Gibbons Ch.J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology.* 2019;19(1). https://doi.org/10.1186/s12874-019-0681-4
2. Гусев А.В., Гаврилов Д.В., Корсаков И.Н., Серова Л.М., Новицкий Р.Э., Кузнецова Т.Ю. Перспективы использования методов машинного обучения для предсказания сердечно-сосудистых заболеваний. *Врач и информационные технологии.* 2019;(3):41–47.
Gusev A.V., Gavrilov D.V., Korsakov I.N., Serova L.M., Novitsky R.E., Kuznetsova T.Yu. Prospects for the use of machine learning methods for predicting cardiovascular disease. *Vrach i informatsionnye tekhnologii = Medical Doctor and Information Technologies.* 2019;(3):41–47. (In Russ.).
3. Гарри Д.Д., Саакян С.В., Хорошилова-Маслова И.П., Цыганков А.Ю., Никитин О.И., Тарасов Г.Ю. Методы машинного обучения в офтальмологии. Обзор литературы. *Офтальмология.* 2020;17(1):20–31. https://doi.org/10.18008/1816-5095-2020-1-20-31
Garri D.D., Saakyan S.V., Khoroshilova-Maslova I.P., Tsygankov A.Yu., Nikitin O.I., Tarasov G.Yu. Methods of Machine Learning in Ophthalmology: Review. *Oftal'mologiya = Ophthalmology in Russia.* 2020;17(1):20–31. (In Russ.). https://doi.org/10.18008/1816-5095-2020-1-20-31
4. Синотова С.Л., Солодушкин С.И., Плаксина А.Н., Макутина В.А. Интеллектуальная система поддержки принятия врачебных решений для прогнозирования исхода протокола вспомогательных репродуктивных технологий на различных этапах его проведения. *Моделирование, оптимизация и информационные технологии.* 2022;10(2). https://doi.org/10.26102/2310-6018/2022.37.2.009
Sinotova S.L., Solodushkin S.I., Plaksina A.N., Makutina V.A. An intelligent clinical decision support system for predicting the outcome of an assisted reproductive technology protocol at various stages of its implementation. *Modelirovanie, optimizatsiya*

i informatsionnye tekhnologii = Modeling, Optimization and Information Technology. 2022;10(2). (In Russ.). https://doi.org/10.26102/2310-6018/2022.37.2.009

5. Синотова С.Л., Лимановская О.В., Плаксина А.Н., Макутина В.А. Программное приложение для предсказания здоровья ребенка, рожденного при помощи вспомогательных репродуктивных технологий, по анамнезу матери. *Моделирование, оптимизация и информационные технологии*. 2021;9(3). https://doi.org/10.26102/2310-6018/2021.34.3.008
Sinotova S.L., Limanovskaya O.V., Plaksina A.N., Makutina V.A. Software application for predicting the health status of a child born with the use of assisted reproductive technologies, according to the mothers anamnesis. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii = Modeling, Optimization and Information Technology.* 2021;9(3). (In Russ.). https://doi.org/10.26102/2310-6018/2021.34.3.008
6. Гусев А.В., Новицкий Р.Э., Ившин А.А., Алексеев А.А. Машинное обучение на лабораторных данных для прогнозирования заболеваний. *ФАРМАКОЭКОНОМИКА. Современная фармакоэкономика и фармакоэпидемиология*. 2021;14(4):581–592. https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115
Gusev A.V., Novitskiy R.E., Ivshin A.A., Alekseev A.A. Machine learning based on laboratory data for disease prediction. *FARMAKOEKONOMIKA. Sovremennaya farmakoekonomika i farmakoepidemiologiya = FARMAKOEKONOMIKA. Modern Pharmacoeconomics and Pharmacoepidemiology.* 2021;14(4):581–592. (In Russ.). https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115
7. Жмудяк М.Л., Повалихин А.Н., Стребуков А.В., Жмудяк А.Л., Устинов Г.Г. Автоматизированная система медицинской диагностики заболеваний с учетом их динамики. *Ползуновский вестник*. 2006;(1):95–106.
Zhmudyak M.L., Povalikhin A.N., Strebukov A.V., Zhmudyak A.L., Ustinov G.G. Avtomatizirovannaya sistema meditsinskoi diagnostiki zabolevanii s uchetom ikh dinamiki. *Polzunovskii vestnik = Polzunovskiy vestnik.* 2006;(1):95–106. (In Russ.).
8. Лимановская О.В., Мещанинов В.Н., Гаврилов И.В. Кластеризация пациентов на основе их функциональных, клинических и антропометрических показателей для построения моделей оценки биовозраста. *Моделирование, оптимизация и информационные технологии.* 2023;11(2). https://doi.org/10.26102/2310-6018/2023.41.2.011
Limanovskaya O.V., Meshchaninov V.N., Gavrilov I.V. Clustering of patients based on their functional, clinical and anthropometric indicators for the construction of models for assessing bio-age. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii = Modeling, Optimization and Information Technology.* 2023;11(2). (In Russ.). https://doi.org/10.26102/2310-6018/2023.41.2.011
9. Вьюгин В.В. *Математические основы машинного обучения и прогнозирования*. Москва: МЦМНО; 2013. 304 с.
V'yugin V.V. *Matematicheskie osnovy mashinnogo obucheniya i prognozirovaniya.* Moscow: MTsMNO; 2013. 304 p. (In Russ.).
10. Кобзарь А.И. *Прикладная математическая статистика.* Москва: ФИЗМАТЛИТ; 2006. 816 с.
Kobzar' A.I. *Prikladnaya matematicheskaya statistika.* Moscow: FIZMATLIT; 2006. 816 p. (In Russ.).
11. Littlestone N., Warmuth M.K. The Weighted Majority Algorithm. *Information and Computation*. 1994;108(2):212–261. https://doi.org/10.1006/inco.1994.1009

ИНФОРМАЦИЯОБАВТОРАХ / INFORMATION ABOUT THE AUTHORS

Лимановская Оксана Викторовна, кандидат химических наук, старший научный сотрудник лаборатории антивозрастных технологий Института медицинских клеточных технологий, Екатеринбург, Российская Федерация.
e-mail: limanovskaya@mail.ru
ORCID: 0000-0002-2084-3916

Oksana V. Limanovskaya, Candidate of Chemical Sciences, Senior Researcher, Laboratory of Anti-Aging Technologies, Institute of Medical Cell Technologies, Yekaterinburg, the Russian Federation.

Гаврилов Илья Валерьевич, кандидат биологических наук, доцент кафедры биохимии Уральского государственного медицинского университета Министерства здравоохранения Российской Федерации, старший научный сотрудник лаборатории антивозрастных технологий Института медицинских клеточных технологий, Екатеринбург, Российская Федерация.
e-mail: iliagavrilov18@yandex.ru
ORCID: 0000-0003-0806-1177

Iliya V. Gavrilov, Candidate of Biological Sciences, Associate Professor at the Department of Biochemistry of Ural State Medical University of the Ministry of Health of the Russian Federation, Senior Researcher, Laboratory of Anti-Aging Technologies, Institute of Medical Cell Technologies, Yekaterinburg, the Russian Federation.

Мещанинов Виктор Николаевич, доктор медицинских наук, профессор, заведующий кафедрой биохимии Уральского государственного медицинский университета Министерства здравоохранения Российской Федерации, заведующий лабораторией антивозрастных технологий Института медицинских клеточных технологий, Екатеринбург, Российская Федерация.
e-mail: mv-02@yandex.ru
ORCID: 0000-0001-7928-2503

Viktor N. Meshchaninov, Doctor of Medical Sciences, Professor, Head of the Department of Biochemistry, Ural State Medical University of the Ministry of Health of the Russian Federation, Head of the Laboratory of Anti-Aging Technologies, Institute of Medical Cell Technologies, Center providing specialized types of medical care, Yekaterinburg, the Russian Federation.

Лисовенко Антон Сергеевич, аспирант кафедры интеллектуальных информационных технологий института фундаментального образования Уральского федерального университета имени первого Президента России Б.Н. Ельцина, Екатеринбург, Российская Федерация.
e-mail: anton.lisovenko.researcher@mail.ru
ORCID: 0000-0001-9127-0820

Anton S. Lisovenko, Postgraduate Student, the Department of Intellectual Information Technologies, Institute of Fundamental Education, Ural Federal University named after the first President of Russia B.N. Yeltsin, Yekaterinburg, the Russian Federation.

Статья поступила в редакцию 24.05.2024; одобрена после рецензирования 10.06.2024; принята к публикации 14.06.2024.

The article was submitted 24.05.2024; approved after reviewing 10.06.2024; accepted for publication 14.06.2024.