


Covid-19



Калькулятор прогнозирования летального исхода пациентов с установленным диагнозом COVID-19

**Корсаков И.Н., Каронова Т.Л., Конради А.О., Рубин А.Д., Курапеев
Д.И., Черникова А.Т., Михайлова А.А., Шляхто Е.В.**

Аннотация

- Целью исследования было проведение системного анализа факторов, влияющих на течение инфекционного заболевания у пациентов с диагностированным COVID-19, госпитализированных в стационар.
- Исследование представляет минимальный риск для людей, поскольку сбор данных был пассивным и не представлял угрозы для участвующих субъектов. Все данные о пациентах хранились в соответствии с Федеральным законом «О персональных данных».
- Пациенты были госпитализированы в ФГБУ «НМИЦ им. В. А. Алмазова» в период с 13 мая 2020 года по конец августа 2021 года.
- В общей сложности были проанализированы данные 4071 пациентов, которые были госпитализированы в течение периода исследования.
- В настоящем исследовании предложен подход ML для прогнозирования летального исхода у пациентов с установленным диагнозом COVID-19 на основе анамнеза пациента и клинических, лабораторных и инструментальных данных, полученных в первые 72 часа нахождения пациента в стационаре.



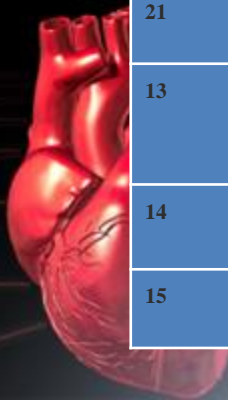
Значимые признаки, полученные в первые 72 часа госпитализации, пациентов для создания алгоритма машинного обучения

	Параметр	Ед.изм.	Количество	Среднее	Отклонение	Ошибка	Доверительный интервал	
1	Возраст	лет	3114.0	62.02	14.33	0.26	61.52	62.52
2	ИМТ	кг/м ²	3114.0	29.46	5.82	0.1	29.25	29.66
3	Ферритин	нг/л	3086.0	658.96	596.22	10.73	637.92	680.01
4	СРБ	мг/л	705.0	67.47	65.21	2.46	62.65	72.3
5	Протромбиновое время	сек.	2981.0	13.95	6.68	0.12	13.71	14.19
6	Гемоглобин	г/л	3113.0	133.77	18.11	0.32	133.13	134.41
7	Лейкоциты	10 ⁹ /л	3113.0	7.79	8.85	0.16	7.48	8.1
8	Нейтрофилы	10 ⁹ /л	2455.0	5.89	4.1	0.08	5.73	6.05
9	Лимфоциты	10 ⁹ /л	3099.0	1.46	6.2	0.11	1.24	1.68
10	Соотношение нейтрофилы/лимфоциты	число	2455.0	6.4	7.74	0.16	6.09	6.7
11	Тромбоциты	10 ⁹ /л	3113.0	221.09	99.92	1.79	217.58	224.6
12	Общий белок	г/л	1352.0	67.92	8.63	0.23	67.46	68.38



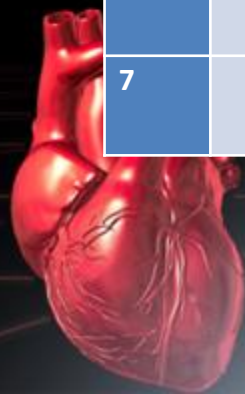
продолжение

	Параметр	Ед.и зм.	Количество	Среднее	Отклонение	Ошибка	Доверительный интервал	
13	Креатинин	мкмол ь/л	3110.0	89.12	52.57	0.94	87.27	90.97
14	Мочевина	ммоль /л	3078.0	6.4	4.19	0.08	6.25	6.55
15	ЧДД	/мин.	3114.0	19.47	5.62	0.1	19.28	19.67
16	Пульсоксиметрия	%	3114.0	95.81	4.29	0.08	95.66	95.96
17	Процент поражения легочной ткани	%	3114.0	36.23	20.27	0.36	35.51	36.94
18	ЧСС	уд/ми н	3114.0	77.75	15.21	0.27	77.22	78.29
19	САД	мм рт.ст	3114.0	125.29	17.4	0.31	124.68	125.9
20	ДАД	мм рт.ст	3114.0	76.33	11.38	0.2	75.93	76.73
21	Температура тела	С°	3114.0	36.94	0.71	0.01	36.92	36.97
13	Креатинин	мкмол ь/л	3110.0	89.12	52.57	0.94	87.27	90.97
14	Мочевина	ммоль /л	3078.0	6.4	4.19	0.08	6.25	6.55
15	ЧДД	/мин.	3114.0	19.47	5.62	0.1	19.28	19.67



Значимые категориальные признаки, полученные в первые 72 часа госпитализации, пациентов для создания алгоритма машинного обучения

	Параметр	Значение	Описание	Количество	Процент
1	Пол	0	женский	1624	52.15
		1	мужской	1490	47.85
2	Исход заболевания	0	выписан	2915	93.61
		1	умер	199	6.39
3	Заболевания бронхолегочной системы	1	наличие	2591	83.2
		0	отсутствует	523	16.8
4	Заболевания ССЗ	0	отсутствует	1570	50.42
		1	наличие	1544	49.58
5	Заболевания эндокринной системы	0	отсутствует	2736	87.86
		1	наличие	378	12.14
6	Онкологические заболевания	0	отсутствует	3054	98.07
		1	наличие	60	1.93
7	Сахарный диабет	0	отсутствует	2891	92.84
		1	наличие	223	7.16



Восстановление пропущенных значений в наборе данных

- Не использовать пациентов у которых есть пропущенные данные
- Вменение с использованием (среднего / медианного) значения
- Вменение с использованием (наиболее часто встречающихся) или (нулевых / постоянных) значений
- Вменение с использованием k ближайших соседей k-NN
- Вменение с использованием многомерного вменения с помощью цепного уравнения (библиотека MICE)
- Вменение с использованием глубокого обучения (библиотека [Datawig](#))



Создание и отбор признаков



Feature transformation – трансформация данных для повышения точности алгоритма

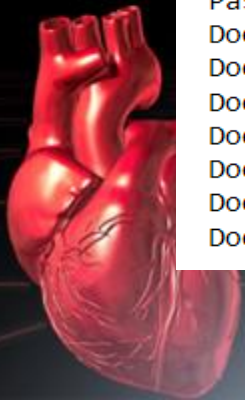
Feature selection – отсечение ненужных признаков

Feature extraction and feature engineering – превращение данных, специфических для предметной области, в понятные для модели векторы



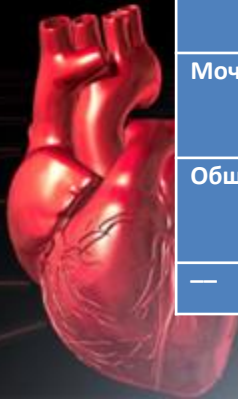
Отбор признаков алгоритм Boruta

Passes the test: Возраст - Ranking: 1
Doesn't pass the test: Рост - Ranking: 16
Doesn't pass the test: bmi - Ranking: 10
Doesn't pass the test: Количество дней в стационаре - Ranking: 3
Doesn't pass the test: Пульсоксиметрия число - Ranking: 7
Passes the test: ЧДД число - Ranking: 1
Passes the test: ЧСС - Ranking: 1
Doesn't pass the test: DAD - Ranking: 7
Passes the test: Процент поражения легочной ткани - Ranking: 1
Doesn't pass the test: D-димер - Ranking: 17
Passes the test: HGB- Гемоглобин - Ranking: 1
Passes the test: PLT- Тромбоциты - Ranking: 1
Passes the test: WBC- Лейкоциты - Ranking: 1
Doesn't pass the test: Глюкоза - Ranking: 2
Passes the test: Д-димер - Ranking: 1
Passes the test: Креатинин - Ranking: 1
Doesn't pass the test: Лимфоциты # (ручн.) - Ranking: 11
Doesn't pass the test: Лимфоциты % (ручн.) - Ranking: 13
Passes the test: Лимфоциты# - Ranking: 1
Passes the test: Лимфоциты% - Ranking: 1
Passes the test: Моноциты% - Ranking: 1
Passes the test: Мочевина - Ranking: 1
Doesn't pass the test: Нейтрофилы# - Ranking: 15
Passes the test: Нейтрофилы% - Ranking: 1
Passes the test: Общий белок - Ranking: 1
Doesn't pass the test: С-реактивный белок (СРБ) - Ranking: 9
Doesn't pass the test: Скорость оседания эритроцитов - Ranking: 14
Doesn't pass the test: Тропонин - Ranking: 5
Doesn't pass the test: Ферритин - Ranking: 5
Doesn't pass the test: Фибриноген - Ranking: 8
Doesn't pass the test: NL - Ranking: 3
Doesn't pass the test: Хронические заболевания сердечно-сосудистой системы_1 - Ranking: 12



Статистические данные полного датасета

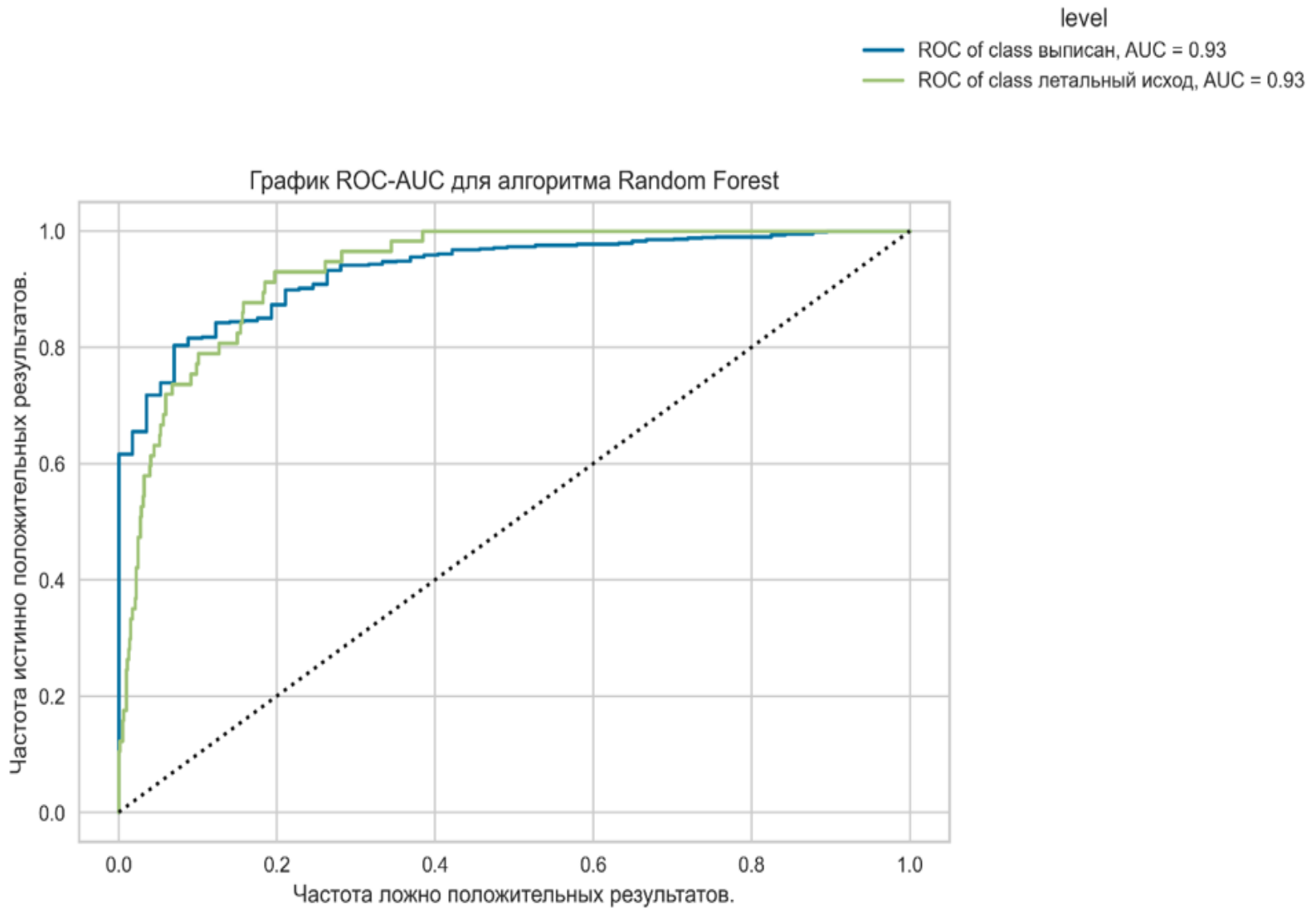
Параметры	count	mean	std	min	25%	50%	75%	max
Возраст, лет	2416.0	61.75	14.26	19.0	52.0	63.0	72.0	95.0
Наличие ССЗ, %	2416.0	0.48	0.5	0.0	0.0	0.0	1.0	1.0
Процент поражения легочной ткани, %	2416.0	35.88	19.98	0.0	20.0	35.0	50.0	100.0
Пульсоксиметрия, %	2416.0	95.87	4.19	6.0	95.0	97.0	98.0	100.0
Тромбоциты, 10 ⁹ /л	2415.0	221.61	99.05	6.0	153.0	201.0	267.5	710.0
СРБ, мг /л	1516.0	56.49	70.6	0.0	5.27	29.68	86.34	445.52
ЧДД, /мин	2416.0	19.43	4.85	0.0	18.0	19.0	20.0	97.0
ЧСС, уд/мин	2416.0	77.39	14.97	37.0	67.0	76.0	87.0	176.0
Креатинин, мкмоль/л	2415.0	88.12	53.68	20.0	67.0	80.0	97.0	1916.0
Мочевина, ммоль/л	2391.0	6.33	4.16	1.1	4.2	5.39	7.3	104.0
Общий белок, г/л	986.0	68.07	8.76	26.2	64.0	69.54	73.65	91.4
—	2416.0	0.06	0.24	0.0	0.0	0.0	0.0	1.0



Результаты машинного обучения (библиотека Rycaret [scikit-learn, CatBoost, XGBoost...])

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Specificity
ada	Ada Boost Classifier	0.0472	0.8863	1.0000	0.0472	0.0901	0.0000	0.0000	0.0000
dummy	Dummy Classifier	0.0472	0.5000	1.0000	0.0472	0.0901	0.0000	0.0000	0.0000
lr	Logistic Regression	0.6491	0.8884	0.9478	0.1140	0.2034	0.1301	0.2528	0.6344
lda	Linear Discriminant Analysis	0.6537	0.8881	0.9367	0.1145	0.2040	0.1308	0.2511	0.6398
nb	Naive Bayes	0.7434	0.8637	0.8256	0.1355	0.2327	0.1650	0.2646	0.7394
rf	Random Forest Classifier	0.8621	0.9081	0.8156	0.2308	0.3589	0.3080	0.3866	0.8645
svm	SVM - Linear Kernel	0.6908	0.0000	0.7611	0.1602	0.2353	0.1699	0.2449	0.6876
qda	Quadratic Discriminant Analysis	0.7382	0.8199	0.7589	0.1250	0.2142	0.1451	0.2327	0.7372
ridge	Ridge Classifier	0.8372	0.0000	0.7500	0.1913	0.3044	0.2477	0.3230	0.8417
gbc	Gradient Boosting Classifier	0.8792	0.9049	0.7478	0.2475	0.3710	0.3229	0.3842	0.8858
et	Extra Trees Classifier	0.8300	0.8811	0.7044	0.1768	0.2819	0.2233	0.2934	0.8362
knn	K Neighbors Classifier	0.7149	0.6916	0.6167	0.0974	0.1679	0.0944	0.1555	0.7198
catboost	CatBoost Classifier	0.9057	0.8957	0.6156	0.2781	0.3816	0.3387	0.3710	0.9200
xgboost	Extreme Gradient Boosting	0.9274	0.9042	0.5044	0.3237	0.3914	0.3549	0.3665	0.9483
lightgbm	Light Gradient Boosting Machine	0.9290	0.9052	0.4489	0.3283	0.3760	0.3398	0.3461	0.9527
dt	Decision Tree Classifier	0.9077	0.6549	0.3756	0.2145	0.2709	0.2254	0.2369	0.9342

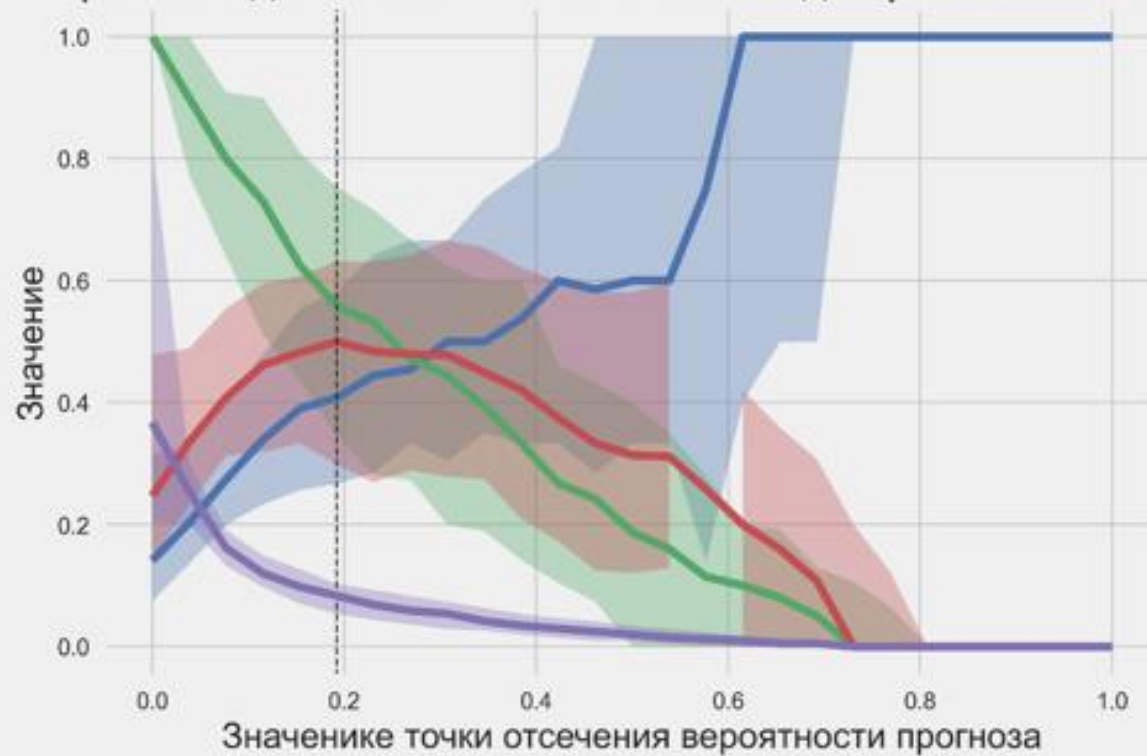
Метрики модели



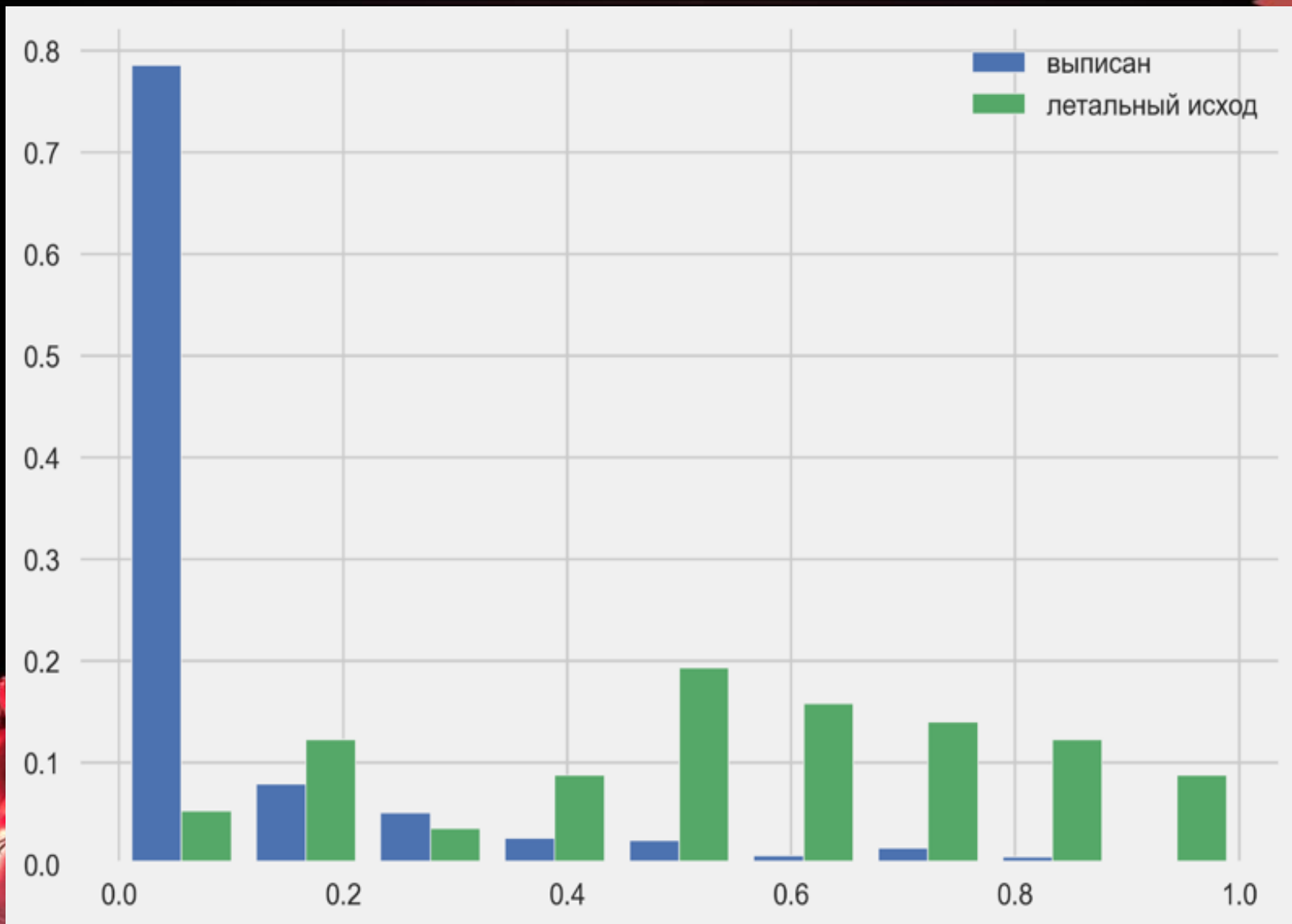
Метрики модели

- Легенда
- precision
 - recall
 - f_1
 - $t_r = 0.19$
 - queue rate

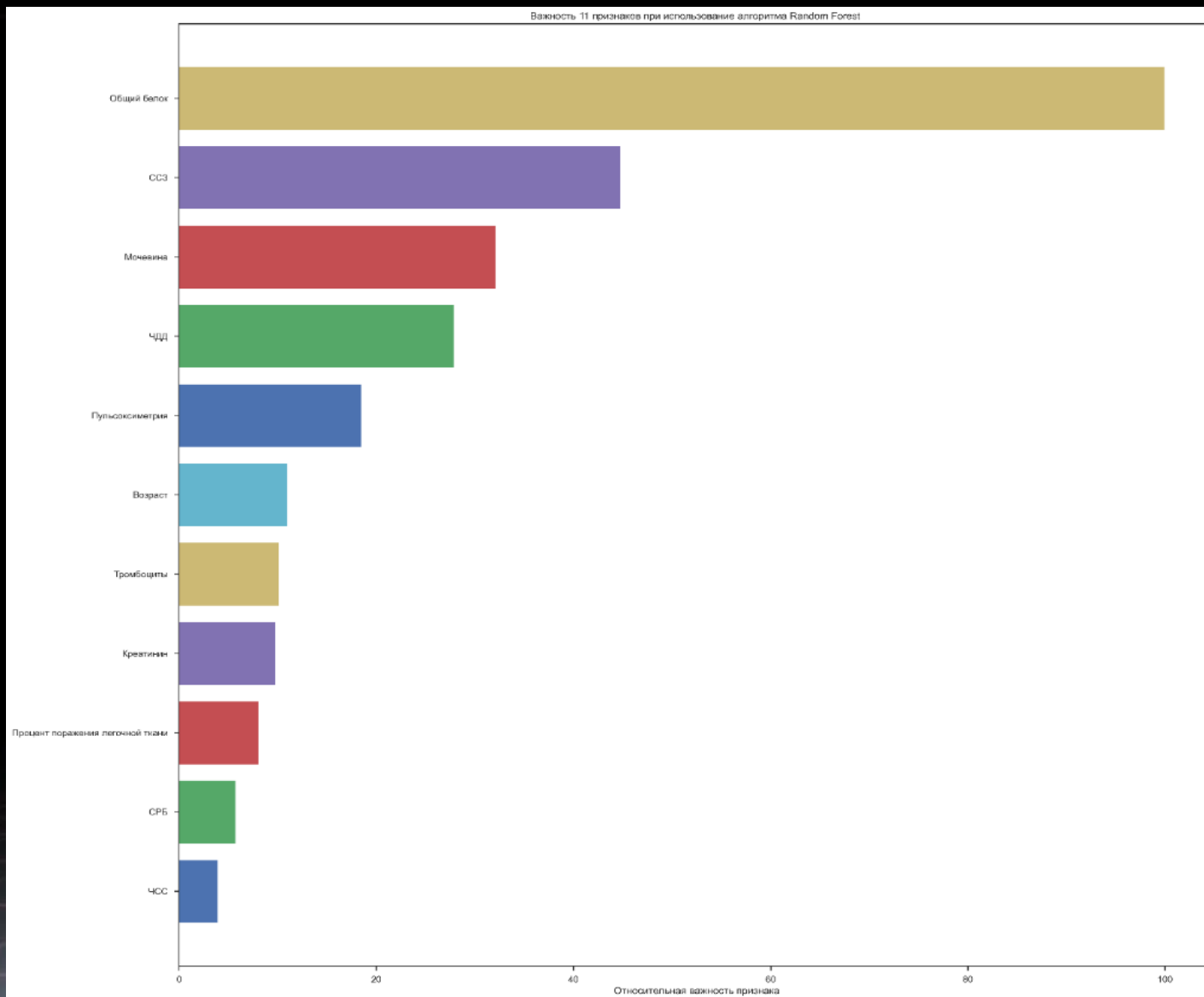
График Метрик метода Random Forest Classifier для разных «точек отсечения»



Распределение вероятности летального исхода (нормированное)



Важность признаков



Библиотека SHAP: влияние признаков на результат модели

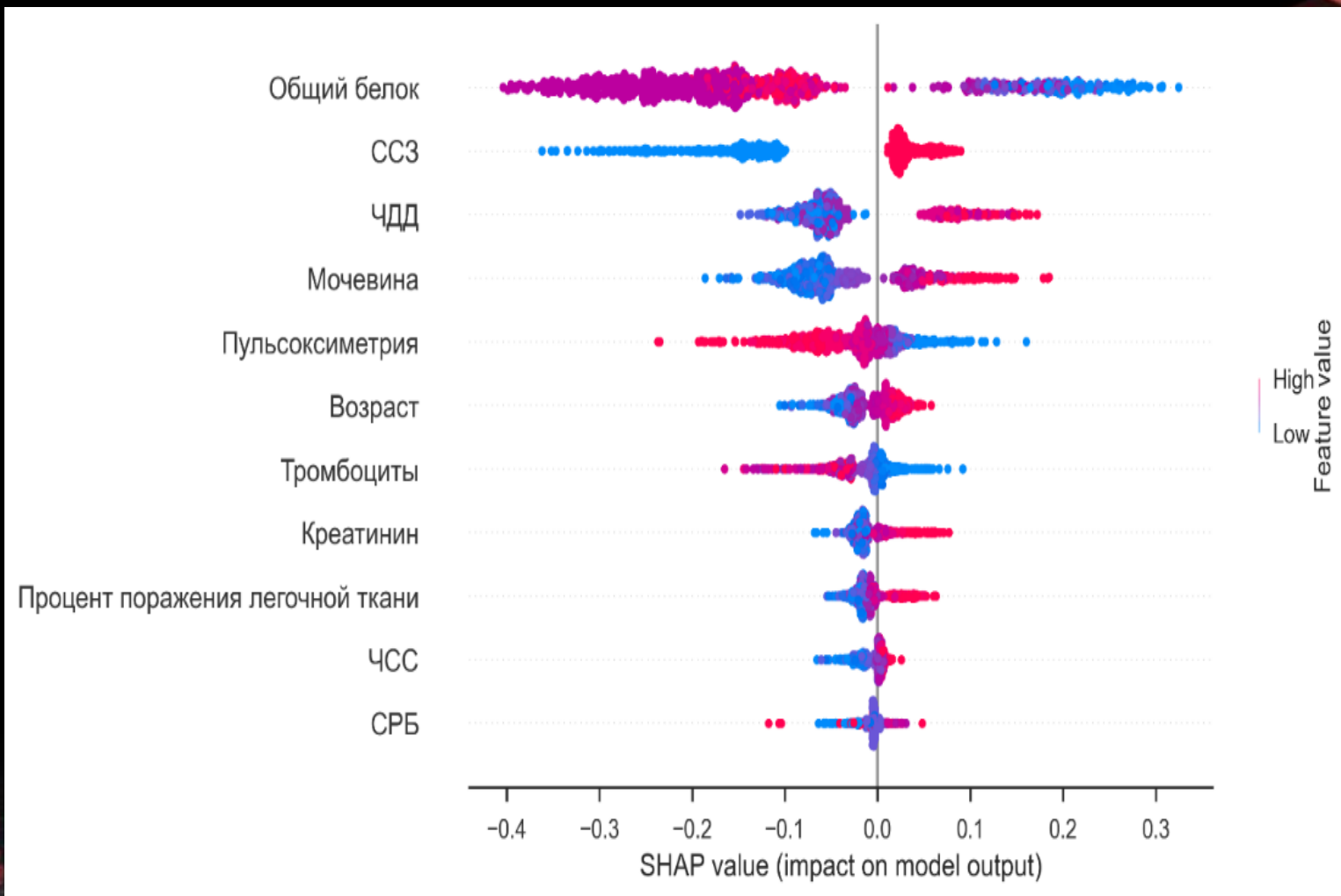


График «Waterfall plot» (SHAP) для пациента 15

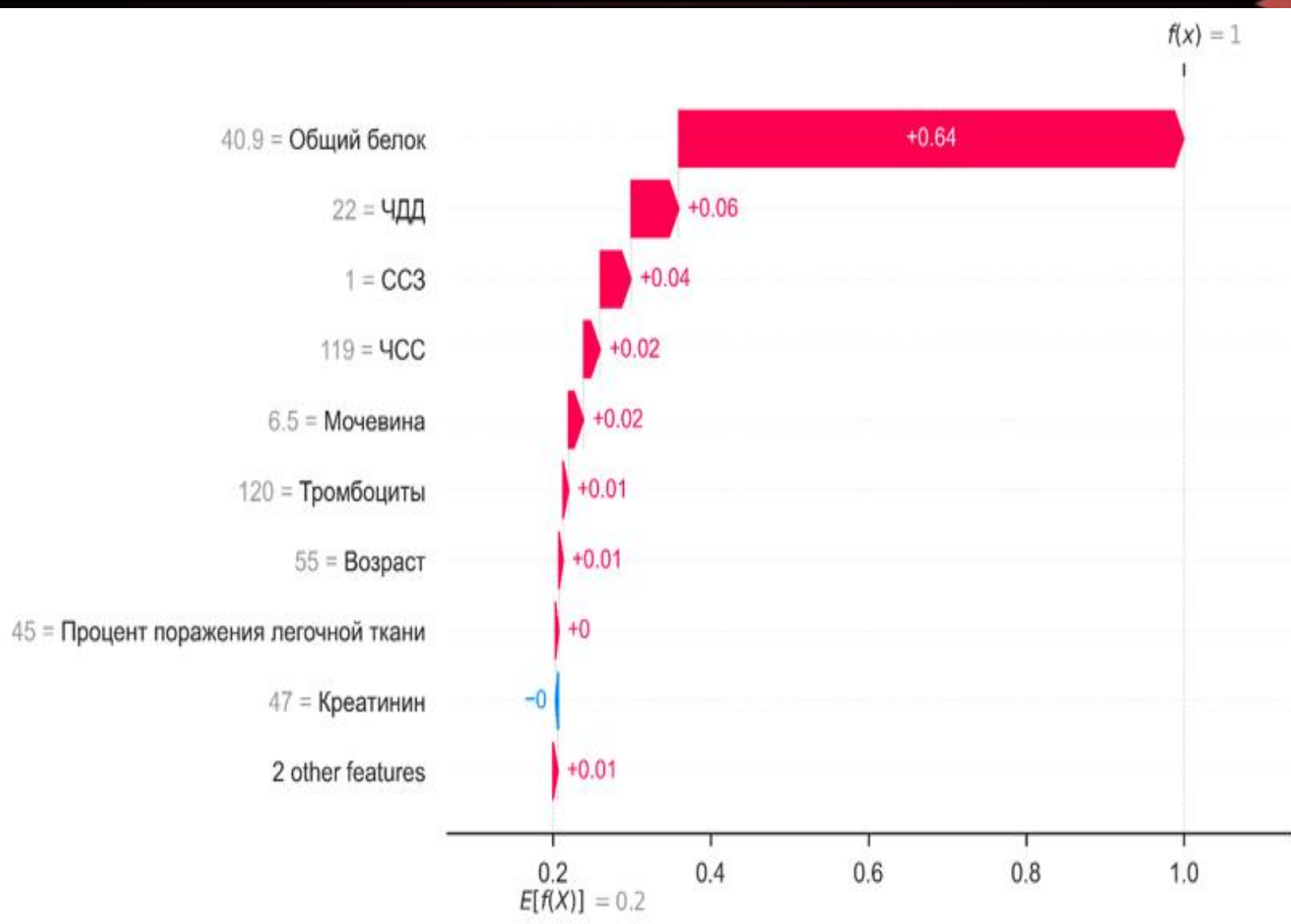
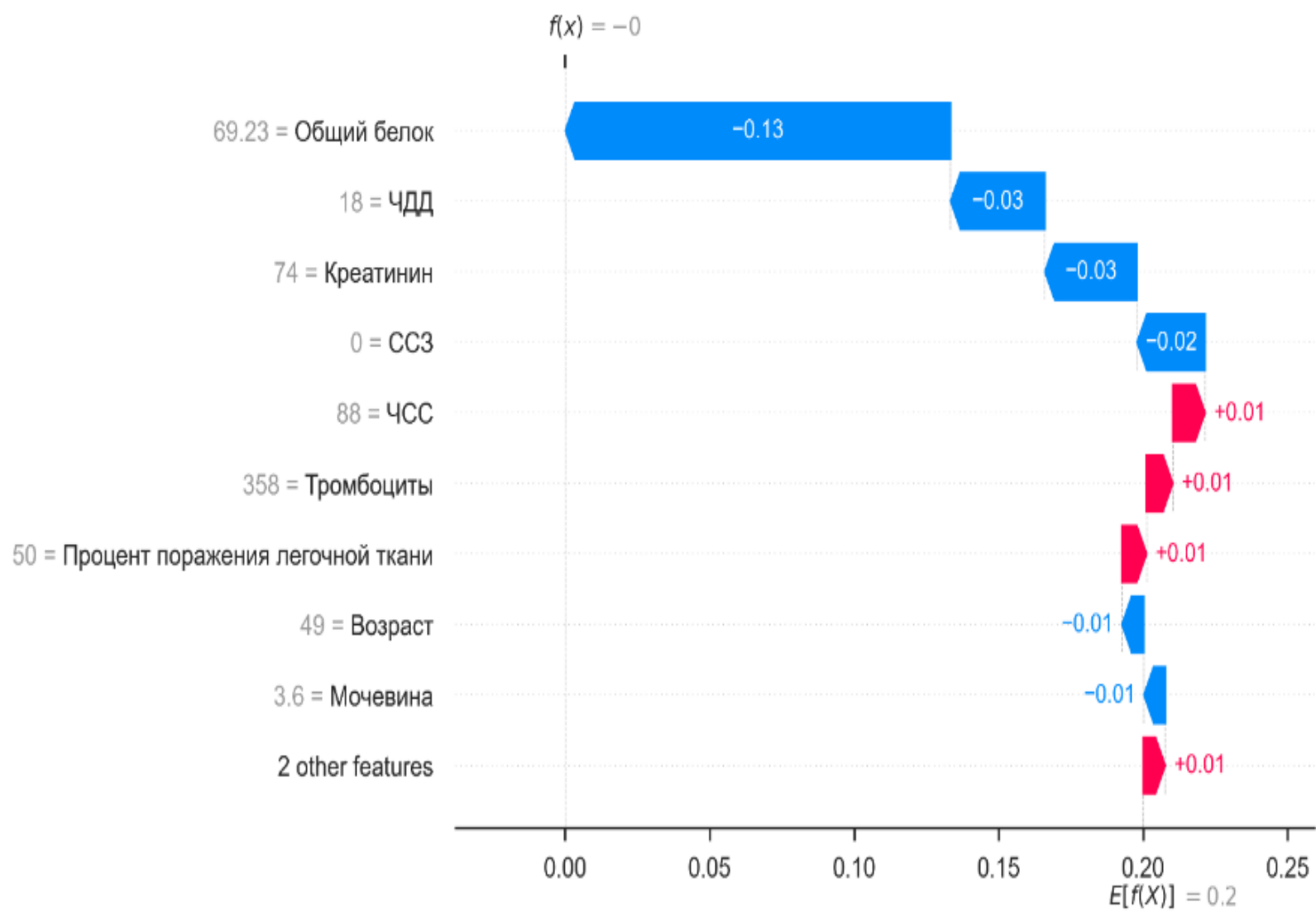
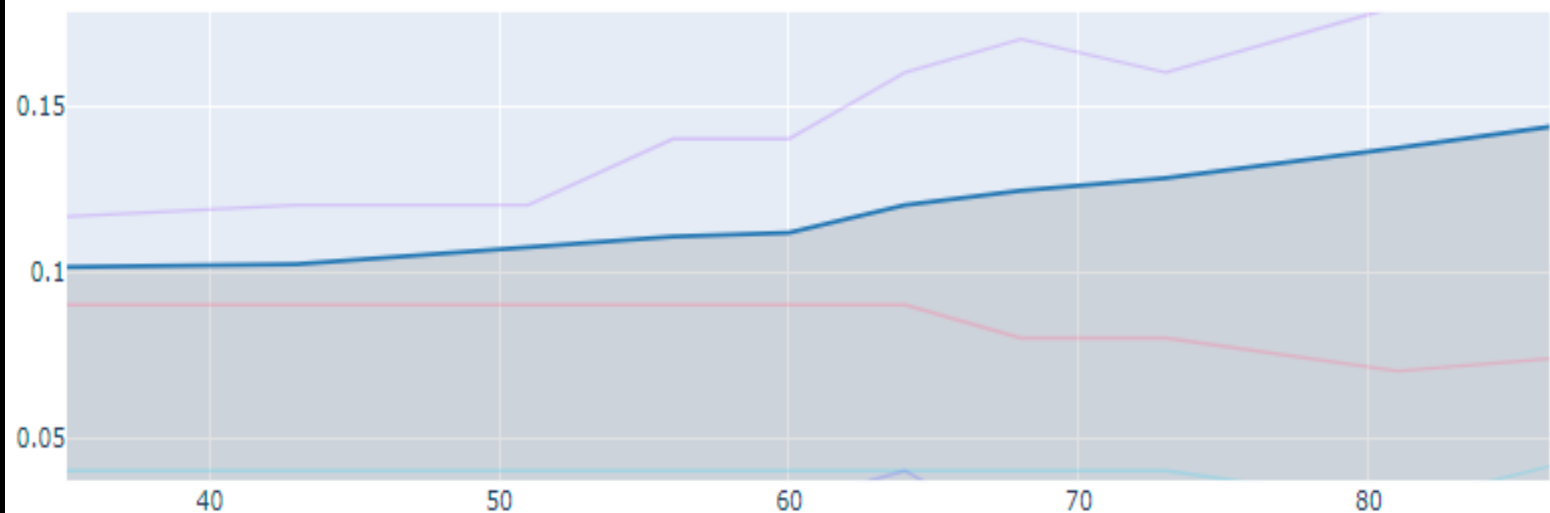


График «Waterfall plot» (SHAP) для пациента 12



Вклад признака «Возраст» в конечную модель и плотность распределения признака.

Возраст



Внешняя валидация

- На тестовых данных

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Specificity
0	Random Forest Classifier	0.8540	0.9317	0.8772	0.2941	0.4405	0.3797	0.4553	0.8524

- На не виденных данных

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Specificity
0	Random Forest Classifier	0.8562	0.9327	0.9231	0.3636	0.5217	0.4553	0.5241	0.8500

- На данных Медучреждения «Б»

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Specificity
0	Random Forest Classifier	0.7302	0.8077	0.8113	0.5119	0.6277	0.4326	0.4609	0.6985

Новые признаки

Облегченная версия калькулятора (Covid-19 calculator lite)

Входные параметры:

1. Возраст
2. ИМТ
3. Пол
4. Тромбоциты
5. Креатинин
6. СРБ
7. Процент поражения легочной ткани
8. ЧСС
9. ЧДД
10. ССЗ
11. Пульсометрия



Метрики для нового калькулятора

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Specificity
Fold								
0	0.8732	0.9477	0.8000	0.2500	0.3810	0.3312	0.4018	0.8769
1	0.8976	0.9492	0.8000	0.2963	0.4324	0.3889	0.4475	0.9026
2	0.8634	0.9010	0.8000	0.2353	0.3636	0.3118	0.3861	0.8667
3	0.8537	0.8800	0.7000	0.2059	0.3182	0.2626	0.3252	0.8615
4	0.8585	0.9077	0.8000	0.2286	0.3556	0.3026	0.3787	0.8615
5	0.8537	0.8849	0.8000	0.2222	0.3478	0.2939	0.3716	0.8564
6	0.8829	0.9049	0.8000	0.2667	0.4000	0.3526	0.4188	0.8872
7	0.8829	0.9464	0.8000	0.2667	0.4000	0.3526	0.4188	0.8872
8	0.8488	0.9041	0.5000	0.1613	0.2439	0.1837	0.2205	0.8667
9	0.9167	0.9467	1.0000	0.3462	0.5143	0.4802	0.5621	0.9128
Mean	0.8731	0.9173	0.7800	0.2479	0.3757	0.3260	0.3931	0.8779
Std	0.0208	0.0260	0.1166	0.0481	0.0676	0.0743	0.0825	0.0180

На не виденных данных

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Specificity
0	Random Forest Classifier	0.8636	0.8541	0.7000	0.3182	0.4375	0.3721	0.4097	0.8770

на тестовых данных

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Specificity
0	Random Forest Classifier	0.8684	0.9100	0.8406	0.3648	0.5088	0.4461	0.4982	0.8708

Результаты

- Алгоритм машинного обучения для прогнозирования летального исхода у пациентов с COVID-19 в течение 72 часов госпитализации продемонстрировал высокую чувствительность (0.842) и специфичность (0.846) и превзошел обычно используемую систему оценки раннего предупреждения; алгоритм способен обнаруживать на 11% больше пациентов, чем модифицированный показатель раннего предупреждения NEWS2 ($p < 0,05$) при одновременном снижении ложноположительных результатов. Учитывая серьезные опасения по поводу ограниченных ресурсов, включая аппараты искусственной вентиляции легких, во время пандемии COVID-19, точное прогнозирование пациентов, которым, вероятно, потребуются искусственная вентиляция легких, может помочь дать важные рекомендации в отношении сортировки пациентов и распределения ресурсов среди госпитализированных больных. Кроме того, раннее выявление таких лиц может позволить проводить плановые процедуры вентиляции легких, снижая некоторые известные риски, связанные с экстренной интубацией. Таким образом, этот алгоритм может помочь улучшить оказание медицинской помощи пациентами, снизить смертность больных и свести к минимуму нагрузку на врачей во время пандемии COVID-19 .



Сотрудничество

- Канал в Телеграм - https://t.me/covid_19_calc
- Калькулятор - <http://85.143.200.21/>
- E-mail – Korsakov_IN@almazovcenter.ru



СПАСИБО!