ITM AI
8-9 февраля 2024

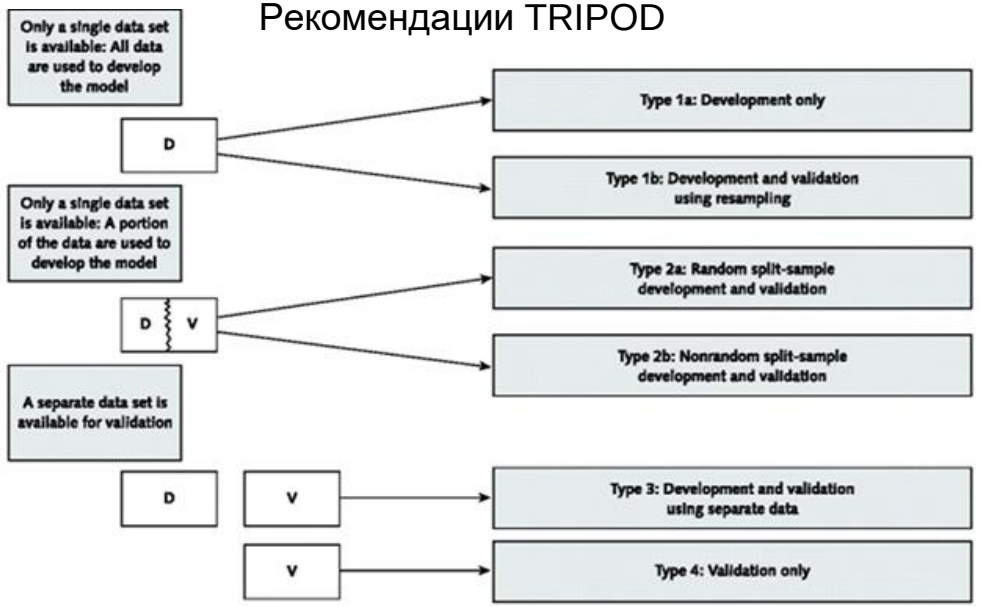
Практическая ценность автоматизации процессов создания и оценки моделей машинного обучения, решающих клинико-практические задачи

Анна Андрейченко, PhD, к.ф.-м.н.
руководитель направления искусственного интеллекта
ООО «К-Скай»

Этапы создания и оценки предиктивных моделей

- Постановка задачи
- Техническое задание
- Моделирование
- Внешняя валидация
- Выбор модели для эксплуатации

Рекомендации TRIPOD



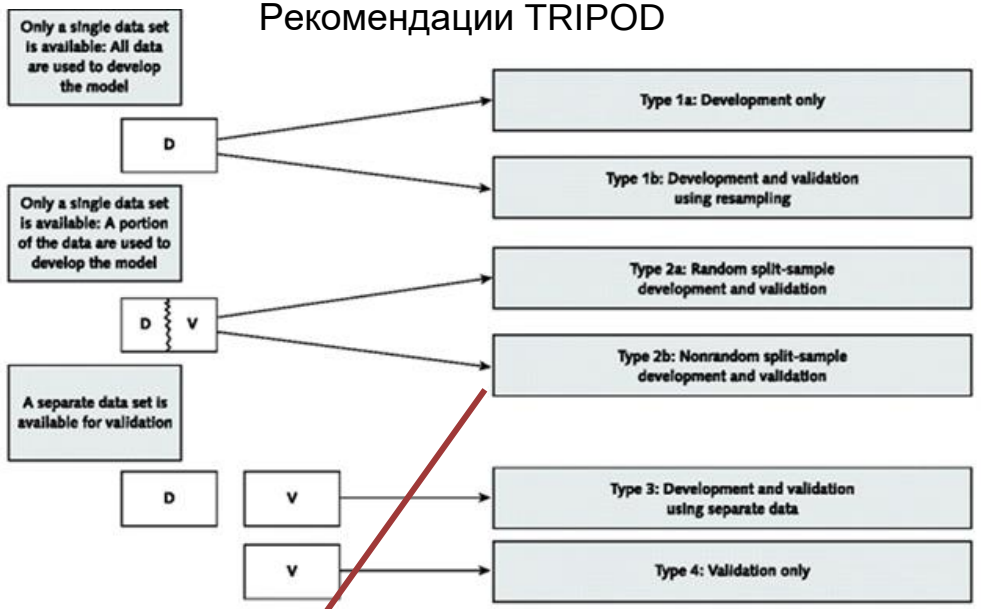
Analysis Type	Description
Type 1a	Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance).
Type 1b	Development of a prediction model using the entire data set, but then using resampling (e.g., bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model. Resampling techniques, generally referred to as "internal validation", are recommended as a prerequisite for prediction model development, particularly if data are limited (6, 14, 15).
Type 2a	The data are randomly split into 2 groups: one to develop the prediction model, and one to evaluate its predictive performance. This design is generally not recommended or better than type 1b, particularly in case of limited data, because it leads to lack of power during model development and validation (14, 15, 16).
Type 2b	The data are nonrandomly split (e.g., by location or time) into 2 groups: one to develop the prediction model and one to evaluate its predictive performance. Type 2b is a stronger design for evaluating model performance than type 2a, because allows for nonrandom variation between the 2 data sets (6, 13, 17).
Type 3	Development of a prediction model using 1 data set and an evaluation of its performance on separate data (e.g., from a different study).
Type 4	The evaluation of the predictive performance of an existing (published) prediction model on separate data (13).

Types 3 and 4 are commonly referred to as "external validation studies." Arguably type 2b is as well, although it may be considered an intermediary between internal and external validation.

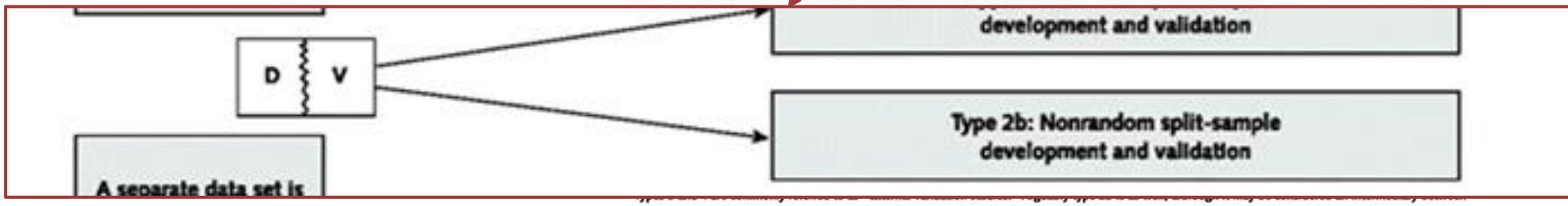
Этапы создания и оценки предиктивных моделей

- Постановка задачи
- Техническое задание
- Моделирование
- Внешняя валидация
- Выбор модели для эксплуатации

Рекомендации TRIPOD



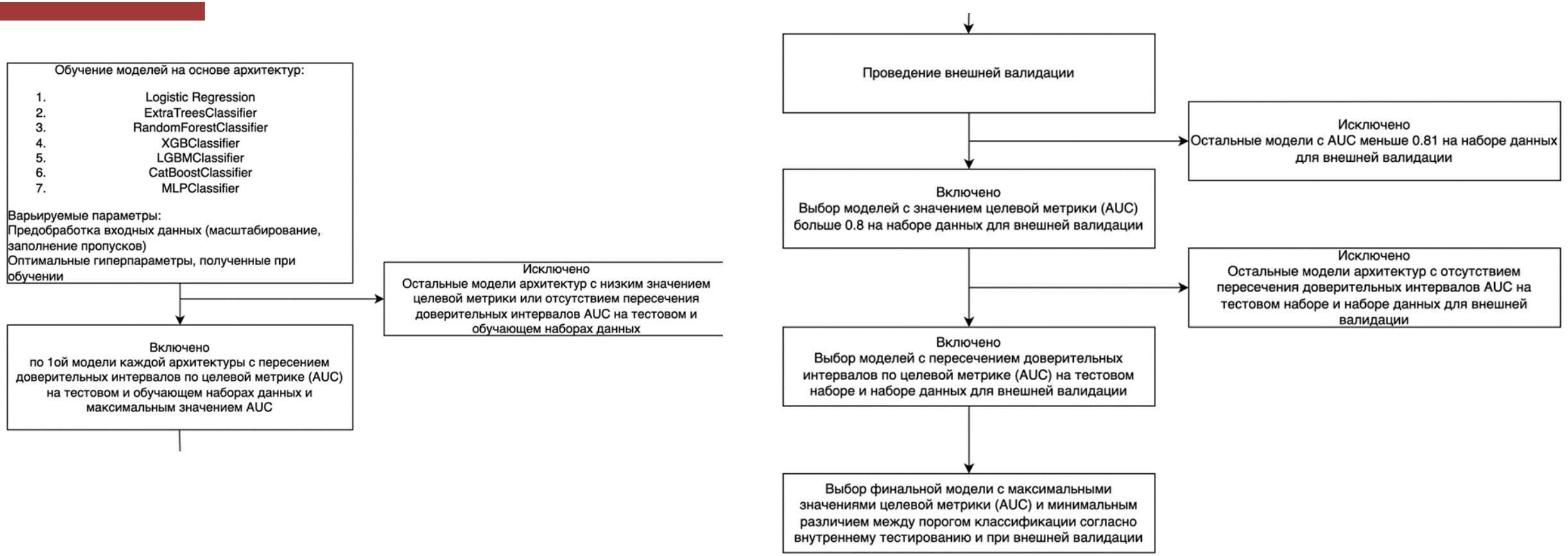
Analysis Type	Description
Type 1a	Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance).
Type 1b	Development of a prediction model using the entire data set, but then using resampling (e.g., bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model. Resampling techniques, generally referred to as "internal validation", are recommended as a prerequisite for prediction model development, particularly if data are limited (6, 14, 15).



internal and external validation.

Схема создания и оценки моделей

Алгоритм выбора итоговой модели



Практические примеры

- Госпитализации пациентов с хроническими заболеваниями в течение ближайших 12 месяцев
- Оценка риска наличия ТЭЛА
- Оценка риска развития преэклампсии у беременных

Прогнозирование госпитализации пациентов с сахарным диабетом в течение 12 месяцев

автоматизация посредством машинного обучения

Код модели: WML.HospitalizationSD

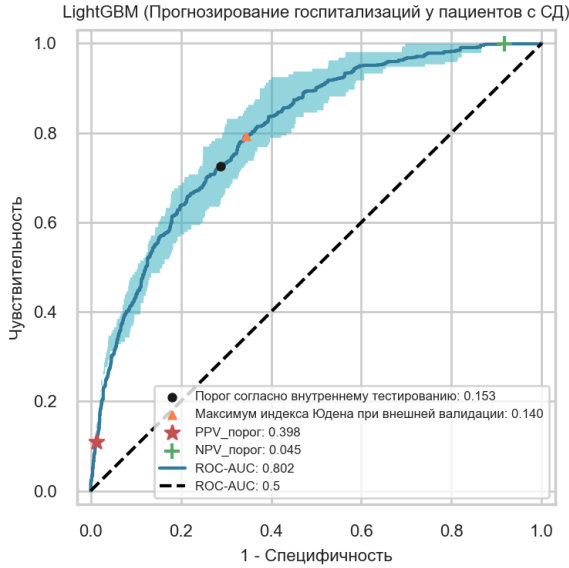
Актуальность

- Распространённость сахарного диабета (СД) как в Российской федерации, так и в мире неуклонно растет в течение нескольких десятилетий
- Популяционный рост и эпидемиологические характеристики СД приводят к колоссальным экономическим расходам и значительному социальному ущербу во всем мире.
- Внедрение многофакторной модели прогнозирования госпитализаций пациентов с СД позволит значительно снизить нагрузку на всю систему здравоохранения за счет раннего вмешательства в течение заболевания у пациентов из группы высокого риска, а также рационального планирования штатной численности и коечного фонда медицинских организаций.

Информация о разработке модели

- **Источник данных:** обезличенные электронные медицинские карты (ЭМК), собранные в платформе Webiomed
- **Набор данных для разработки:** 142 927 записи (19 665 с целевым событием) из них:
  - Обучение: 128 634 (17 698 с целевым событием)
  - Тестирование: 14 293 (1 967 с целевым событием)
- **Набор данных внешней валидации:** 27 134 записей (959 с целевым событием)
- В процессе обучения модели количество входных признаков было сокращено с 118 до 33
- Метод машинного обучения: LightGBM

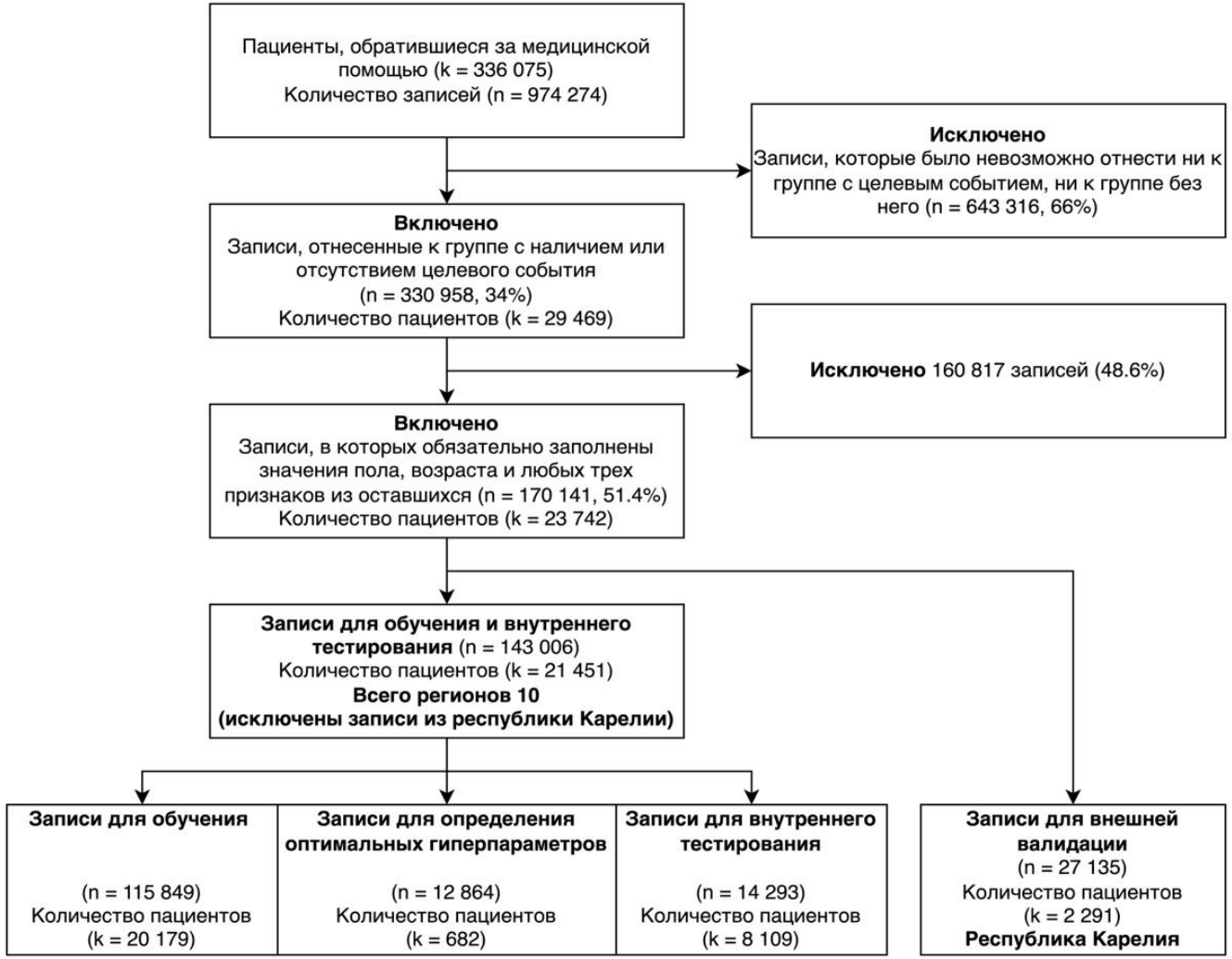
Метрики модели при внешней валидации



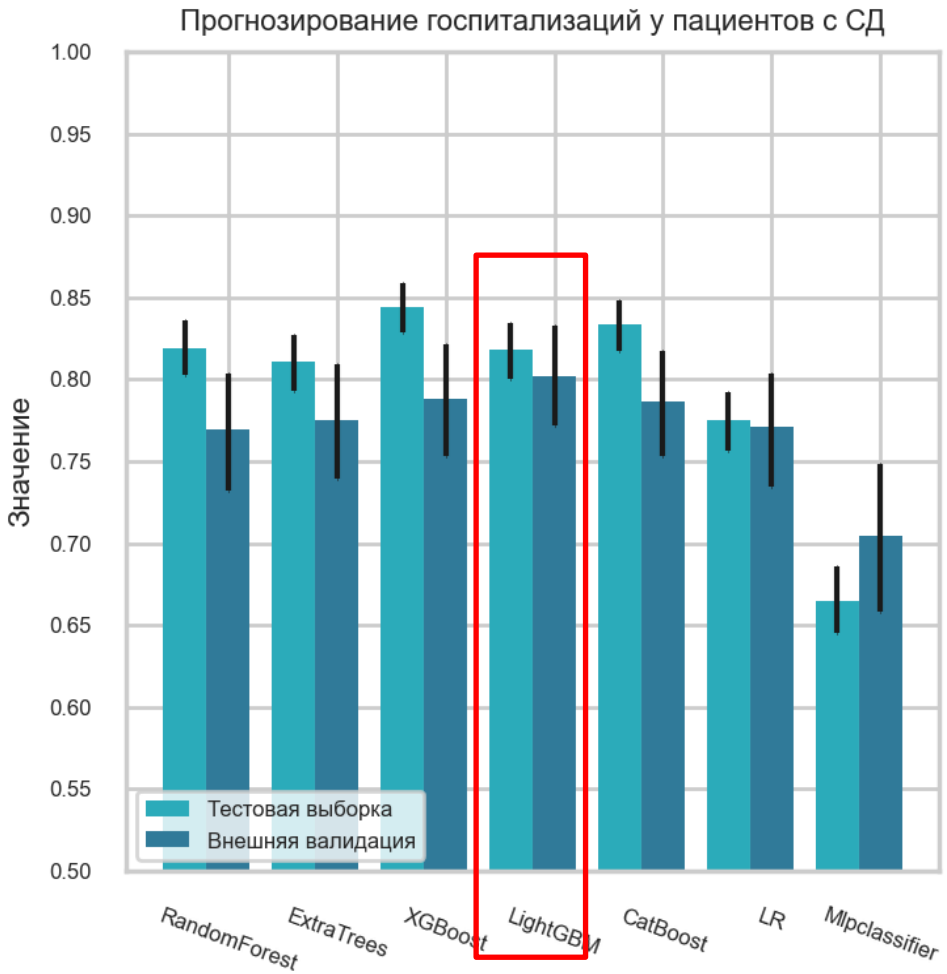
Характеристики и применение модели

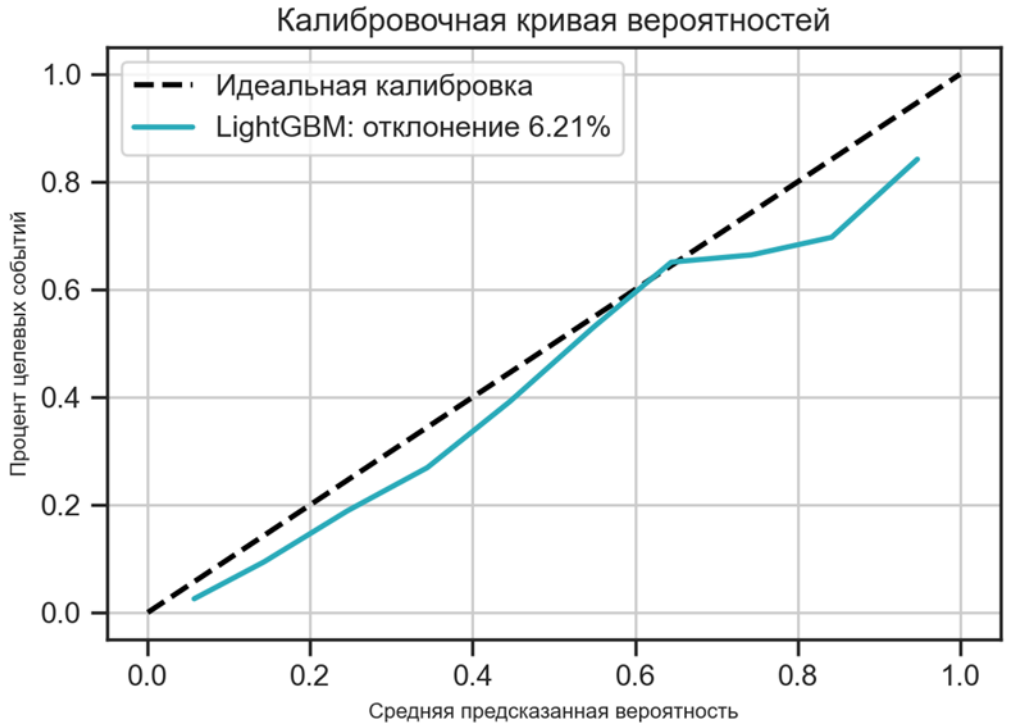
- **Входные данные**: 33 признака, включая антропометрические/лабораторные/инструментальные/анамнестические данные, симптомы
- **Выходные данные**: госпитализации пациента с СД по поводу основного заболевания или его осложнений, 0-100%
- **Значимые признаки:** установленный диагноз диабетической нефропатии, нейропатии, ФП и дислипидемии, жажда, лихорадка, пол, возраст и окружность талии пациента.
- **Метрики качества на наборе данных для внешней валидации при использовании порога №1 (0.045):**
  - Чувствительность: **99.9% (99.4-100%) CI 95%**
  - Прогностическая ценность отрицательного результата (ПЦОР): **100% (99.7-100%) CI 95%**
- **Метрики качества на наборе данных для внешней валидации при использовании порога №2 (0.398):**
  - Специфичность: **98.9% (98.5-99.1%) CI 95%**
  - Прогностическая ценность положительного результата (ПЦПР): **24.9% (15.5 – 35.6%) CI 95%**
- **Назначение и условия применения:**
  - применима для пациентов в возрасте от 18 лет с установленным диагнозом СД
  - обращает внимание врача при значимых факторах риска госпитализации
  - предназначена для поддержки принятия врачебных решений по тактике ведения пациентов с СД

Риск госпитализации пациентов с СД в течение 12 месяцев: набор данных



Риск госпитализации пациентов с СД в течение 12 месяцев: внешняя валидация и калибровка





ITM AI
8-9 февраля 2024

Прогнозирование госпитализации пациентов с гипертонической болезнью в течение 12 месяцев
(автоматизация посредством машинного обучения)
Код модели: WML.HospitalizationCVD.Hypertension

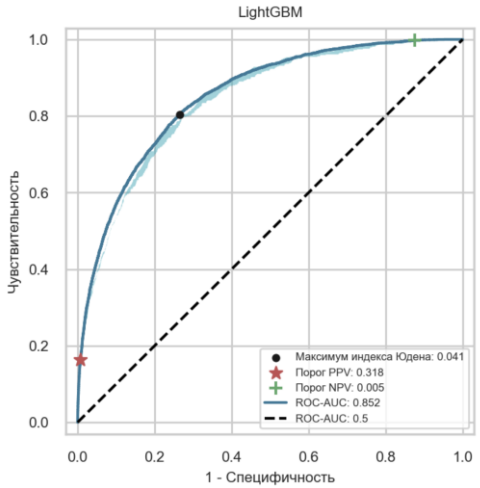
Актуальность

- В РФ высокая распространенность АГ и низкий контроль данного заболевания. 66% госпитализаций больных с гипертонической болезнью можно предотвратить с помощью своевременного лечения на уровне учреждений первичной медицинской помощи.
- РФ: регионы с повышенными уровнями госпитализации по поводу АГ.
- Внедрение многофакторной модели прогнозирования госпитализаций пациентов с АГ позволит значительно снизить нагрузку на систему здравоохранения за счет раннего вмешательства в течение заболевания у пациентов из группы высокого риска, а также рационального планирования штатной численности и коечного фонда.

Информация о разработке модели

- **Источник данных:** обезличенные электронные медицинские карты (ЭМК), собранные в платформе Webiomed
- **Набор данных для разработки:** 1 121 243 записей (49 634 с целевым событием) из них:
  - Обучение: 1 009 119 (44 671 с целевым событием)
  - Тестирование: 107 161 (4 963 с целевым событием)
- **Набор данных внешней валидации:** 112 124 записей (4 963 с целевым событием)
- Метод машинного обучения: LightGBM

Метрика модели при внутреннем тестировании



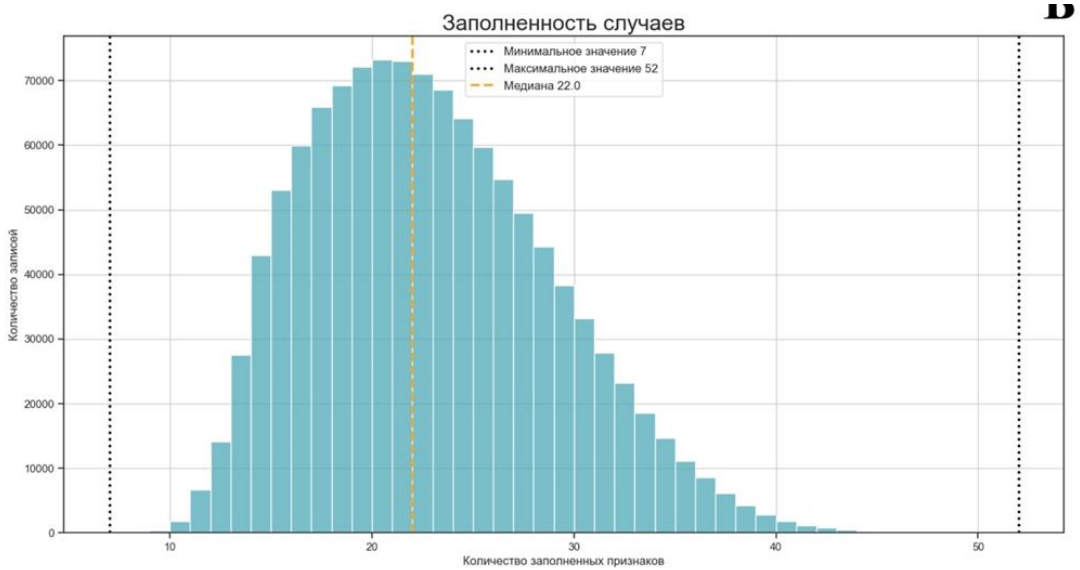
Характеристики и применение модели

- **Входные данные**: 43 признака, включая антропометрические / лабораторные / инструментальные / анамнестические данные, симптомы
- **Выходные данные**: госпитализации пациента с ГБ по поводу основного заболевания или его осложнений, 0-100%
- **Значимые признаки:** фракция выброса левого желудочка, возраст, одышка, пол, количество обращений в поликлинику, цианоз кожи, количество госпитализаций.
- **Метрики качества на наборе данных для внутреннего тестирования при использовании порога №1 (0.005):**
  - Чувствительность: **99.8% (99-100%) CI 95%**
  - Прогностическая ценность положительного результата (ПЦПР): **99.9% (99.6-100%) CI 95%**
- **Метрики качества на наборе данных для внутреннего тестирования при использовании порога №2 (0.318):**
  - Специфичность: **99.2% (99-99.5%) CI 95%**
  - Прогностическая ценность отрицательного результата (ПЦОР): **96.2% (95.7 – 96.8%) CI 95%**
- **Назначение и условия применения:**
  - применима для пациентов в возрасте от 18 лет с установленным диагнозом ГБ
  - обращает внимание врача при значимых факторах риска госпитализации
  - предназначена для поддержки принятия врачебных решений по тактике ведения пациентов с ГБ

Риск госпитализации пациентов с ГБ в течение 12 месяцев: пропуски в признаках и баланс классов

Алгоритм	Заполнение пропусков	Масштабирование	Коррекция дисбаланса	Доля записей из класса 1
Multi-layer Perceptron	Фиксированное значение	Стандартизация	Не проводилась	
LR	Фиксированное значение	Не проводилось	ADASYN	0.1
CatBoost	Не проводилось	Не проводилось	ROS	0.2
XGBoost	Не проводилось	Не проводилось	RUS	0.1
LightGBM	Не проводилось	Не проводилось	Не проводилась	
ExtraTrees	Фиксированное значение	Не проводилось	Не проводилась	
RandomForest	Фиксированное значение	Не проводилось	ROS	0.2

Примечание. LR – Logistic Regression, ADASYN – Adaptive Synthetic Sampling Approach, ROS – Random Oversampling, RUS – Random Undersampling



Риск госпитализации пациентов с ГБ в течение 12 месяцев: метрики

	TN	FP	FN	TP	Площадь под характеристической кривой (AUC)	Доля правильных ответов (Accuracy)	Чувствительность	Специфичность	Прогностическая ценность положительного класса (Precision)	Прогностическая ценность отрицательного класса	LR+	LR-
Порог PPV	106349	812	4151	812	0.852 [0.827 0.875]	0.956 [0.95 0.961]	0.164 [0.116 0.215]	**0.992 [0.99 0.995]**	**0.5 [0.382 0.614]**	0.962 [0.957 0.967]	**21.592 [13.792 33.651]**	0.843 [0.791 0.892]
Порог NPV	13370	93791	12	4951	0.852 [0.824 0.874]	0.163 [0.153 0.173]	**0.998 [0.99 1.]**	0.125 [0.116 0.134]	0.05 [0.044 0.057]	**0.999 [0.996 1.]**	1.14 [1.124 1.153]	**0.019 [0. 0.082]**

Прогнозирование госпитализации пациентов с хронической ишемической болезнью сердца в течение 12 месяцев (автоматизация посредством машинного обучения)

Код модели: WML.HospitalizationCVD.CAD

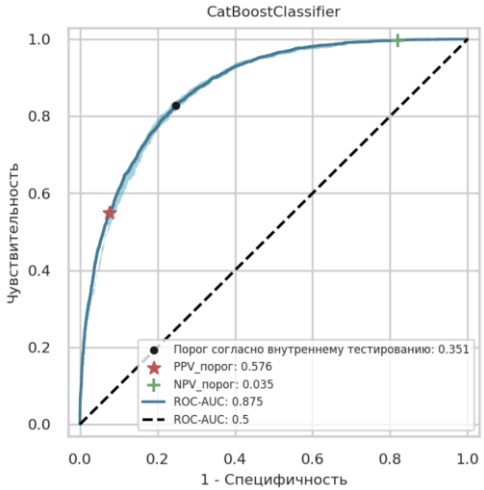
Актуальность

- Распространённость хронической ишемической болезни сердца (ИБС) как в Российской федерации, так и в мире неуклонно растет в течение нескольких десятилетий
- Популяционный рост и эпидемиологические характеристики ИБС приводят к колоссальным экономическим расходам и значительному социальному ущербу во всем мире.
- Внедрение многофакторной модели прогнозирования госпитализаций пациентов с ИБС позволит значительно снизить нагрузку на всю систему здравоохранения за счет раннего вмешательства в течение заболевания у пациентов из группы высокого риска, а также рационального планирования штатной численности и коечного фонда медицинских организаций.

Информация о разработке модели

- **Источник данных:** обезличенные электронные медицинские карты (ЭМК), собранные в платформе Webiomed
- **Набор данных для разработки:** 122 738 записей (43 828 с целевым событием) из них:
  - Обучение: 110 464 (39 445 с целевым событием)
  - Тестирование: 12 274 (4 383 с целевым событием)
- **Набор данных внешней валидации:** 13 135 записей (1 476 с целевым событием)
- Метод машинного обучения: CatBoost

Метрика модели при внутреннем тестировании

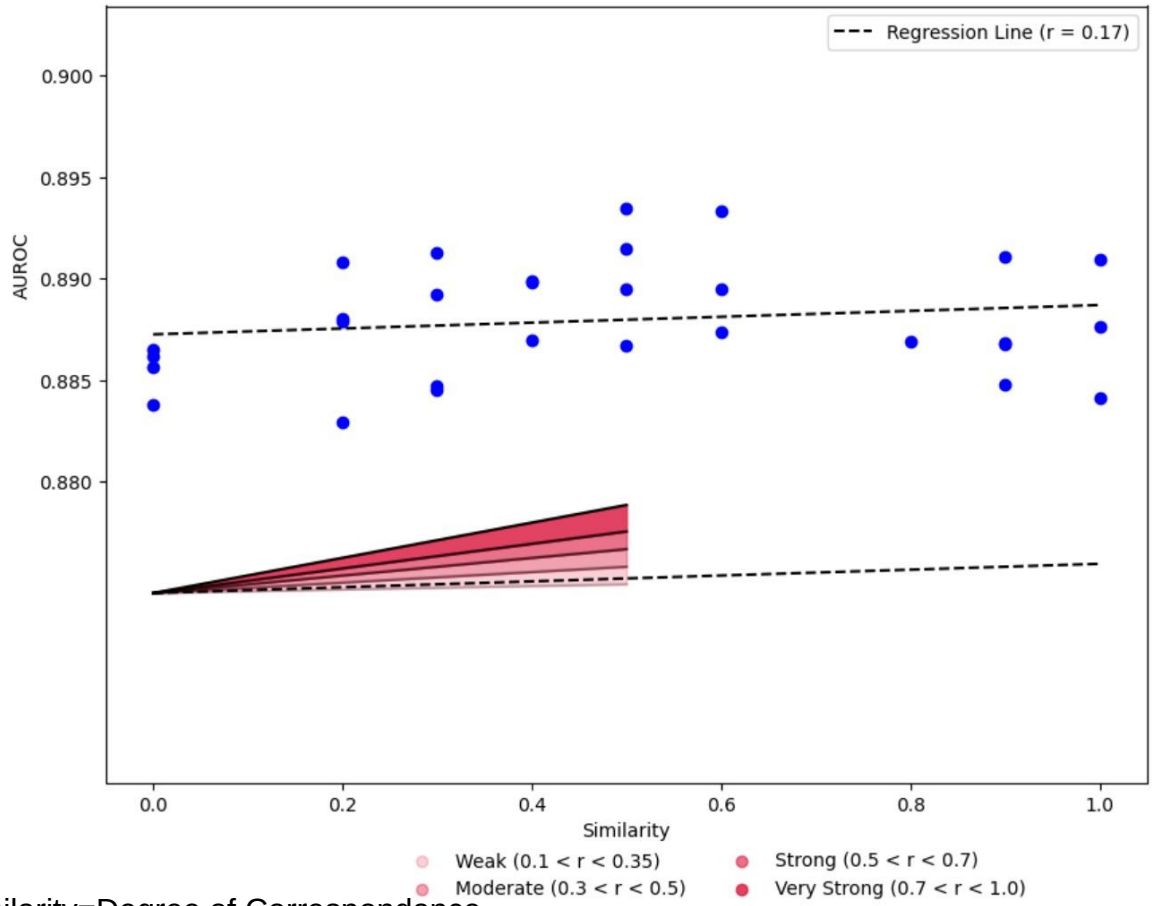


Характеристики и применение модели

- **Входные данные**: 46 признаков, включая антропометрические/лабораторные/инструментальные/анамнестические данные, симптомы
- **Выходные данные**: вероятность госпитализации пациента с ИБС по поводу основного заболевания или его осложнений, 0-100%
- **Значимые признаки:** сердцебиение, фибрилляция предсердий, легочная гипертензия, нерегулярный ритм, одышка, общий холестерин, глюкоза крови, отеки ног.
- **Метрики качества на наборе данных для внутреннего тестирования при использовании порога №1 (0.035):**
  - Чувствительность: **99.7% (99.4-99.9%) CI 95%**
  - Прогностическая ценность отрицательного результата (ПЦОР): **99% (98.1-99.8%) CI 95%**
- **Метрики качества на наборе данных для внутреннего тестирования при использовании порога №2 (0.576):**
  - Специфичность: **92.4% (91.5-93.3%) CI 95%**
  - Прогностическая ценность положительного результата (ПЦПР): **80% (77.9 – 82.3%) CI 95%**
- **Назначение и условия применения:**
  - применима для пациентов в возрасте от 18 лет с установленным диагнозом ИБС
  - обращает внимание врача при значимых факторах риска госпитализации
  - предназначена для поддержки принятия врачебных решений по тактике ведения пациентов с ИБС

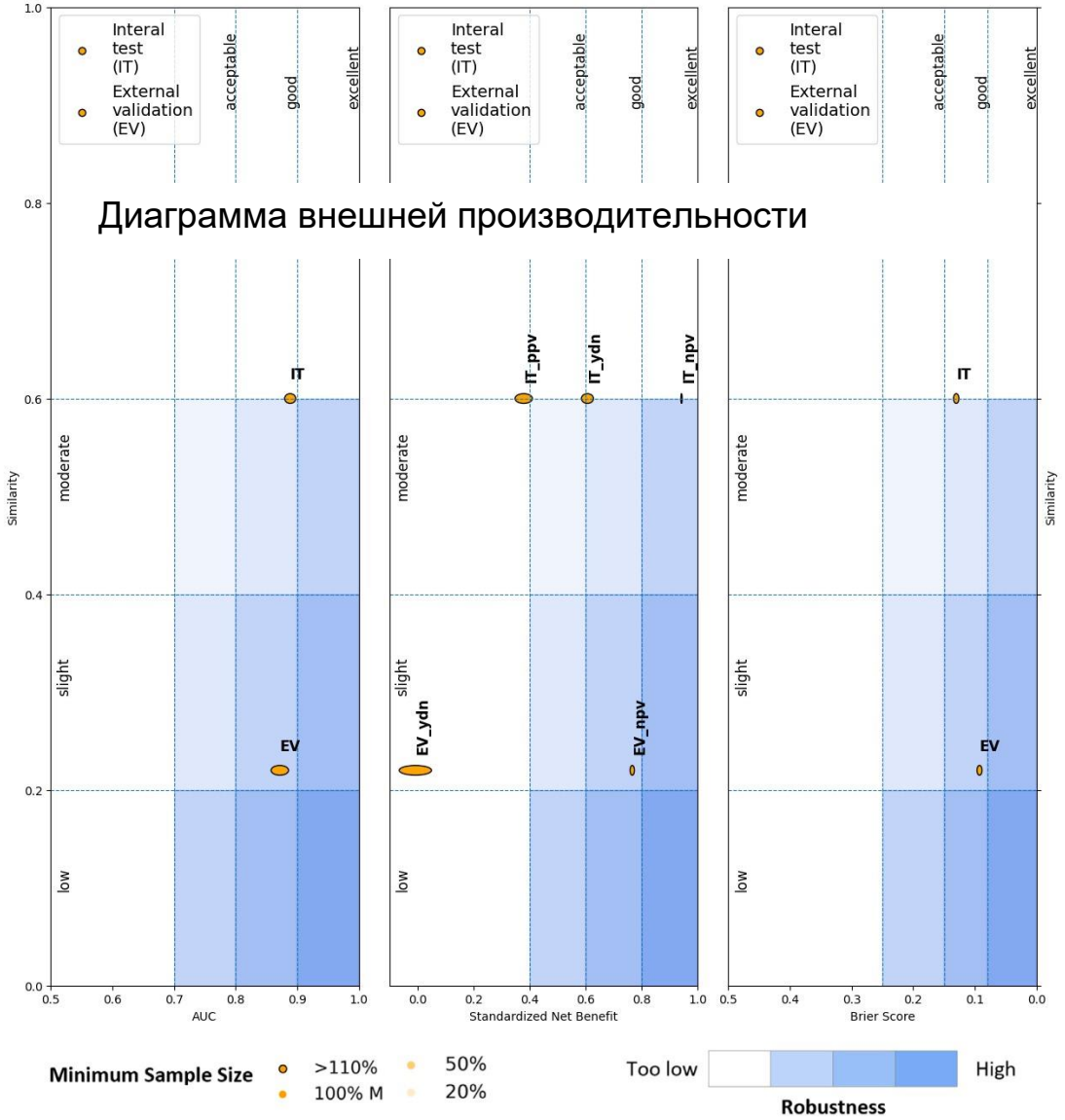
Риск госпитализации пациентов с ИБС в течение 12 месяцев: мета-валидация = размер + схожесть датасетов

Диаграмма потенциальной устойчивости



Similarity=Degree of Correspondance

Диаграмма внешней производительности



Прогнозирование развития преэклампсии (ПЭ) на 7-16 неделе беременности

автоматизация посредством машинного обучения

Код модели: WML.Prognosis.Obstetrics.Preeclampsia.group1

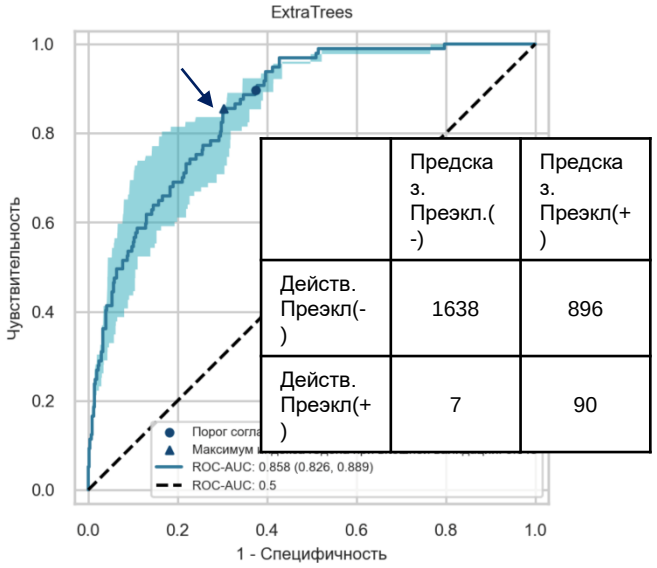
Актуальность

- Распространенность ПЭ 2-8% всех беременных, одна из ведущих причин материнской и перинатальной заболеваемости и смертности в мире
- ПЭ имеет многофакторную природу и является результатом суммарного воздействия различных факторов
- имеется перспектива путем объединения факторов риска для лучшего прогнозирования развития ПЭ и выработки оптимальной тактики ведения беременной

Информация о разработке модели

- Источник данных: обезличенные электронные медицинские карты (ЭМК), собранные в платформе Webiomed
- Набор данных 18461 записи (672 с преэклампсией) из них:
  - Обучение: 16615 (605 с преэклампсией)
  - Тестирование: 1846 (67 с преэклампсией)
- Набор данных внешней валидации: 2631 записей (97 с преэклампсией)
- В процессе обучения модели количество входных признаков было сокращено с 55 до 34
- Метод машинного обучения: ExtraTreesClassifier

Метрики модели при внешней валидации



	Предска з. Преэкл.(-)	Предска з. Преэкл(+)
Действ. Преэкл(-)	1638	896
Действ. Преэкл(+)	7	90

Характеристики и применение модели

- **Входные данные**: 34 признака, включая антропометрические/лабораторные/инструментальные/анамнестические данные, симптомы
- **Выходные данные**: вероятность возникновения преэклампсии, 0-100%
- **Значимые признаки:** отеки, задержка роста плода, ССЗ, ИМТ во время беременности, головная боль, табакокурение, первородящая/нет, вес, неврологические заболевания, дистресс плода
- **Метрики качества на наборе данных для внешней валидации:**
  - Специфичность: **62% (60-64%) CI 95%**
  - Прогностическая ценность положительного результата (ПЦПР): **8% (6 – 10%) CI 95%**
  - Прогностическая ценность отрицательного результата (ПЦОР): **99% (98-99%) CI 95%**
- **Назначение и условия применения:**
  - применима для пациентов в возрасте от 10 до 60 лет на 7-16 неделе беременности
  - обращает внимание врача при значимых прогностических факторах ПЭ
  - предназначена для поддержки принятия врачебных решений по тактике ведения беременных

Прогнозирование развития ранней преэклампсии (ПЭ) на 7-16 неделе беременности

автоматизация посредством машинного обучения

Код модели: WML.Prognosis.Obstetrics.Preeclampsia.group1P

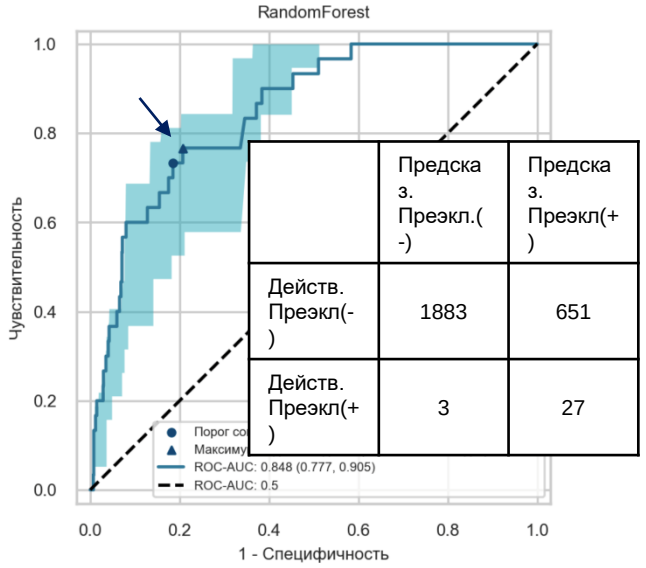
Актуальность

- Распространенность ПЭ 2-8% всех беременных, одна из ведущих причин материнской и перинатальной заболеваемости и смертности в мире
- ПЭ имеет многофакторную природу и является результатом суммарного воздействия различных факторов
- имеется перспектива путем объединения факторов риска для лучшего прогнозирования развития ПЭ и выработки оптимальной тактики ведения беременной

Информация о разработке модели

- Источник данных: обезличенные электронные медицинские карты (ЭМК), собранные в платформе Webiomed
- Набор данных 17952 записей (163 с преэклампсией) из них:
  - Обучение: 16157 (147 с преэклампсией)
  - Тестирование: 1795 (16 с преэклампсией)
- Набор данных внешней валидации: 2564 записей (30 с преэклампсией)
- В процессе обучения модели количество входных признаков было сокращено с 55 до 36
- Метод машинного обучения: RandomForest

Метрики модели при внешней валидации

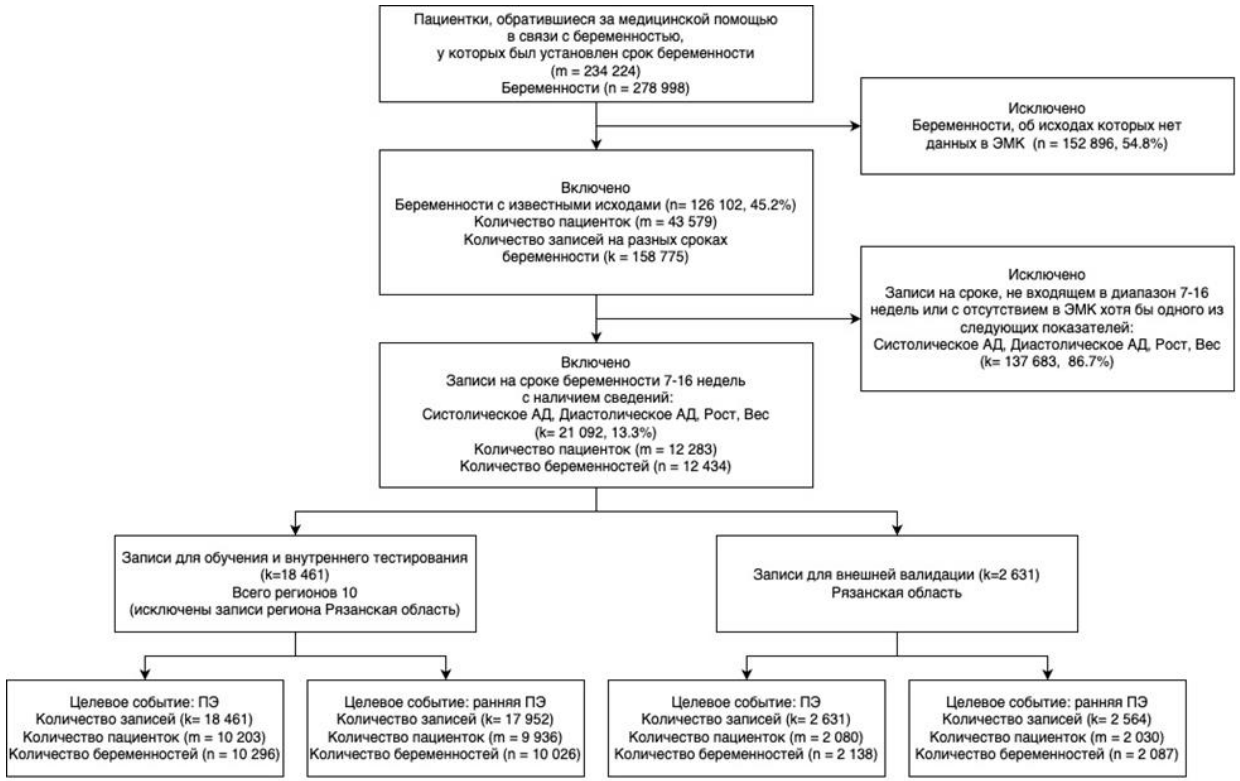


	Предска з. Преэкл.(-)	Предска з. Преэкл(+)
Действ. Преэкл(-)	1883	651
Действ. Преэкл(+)	3	27

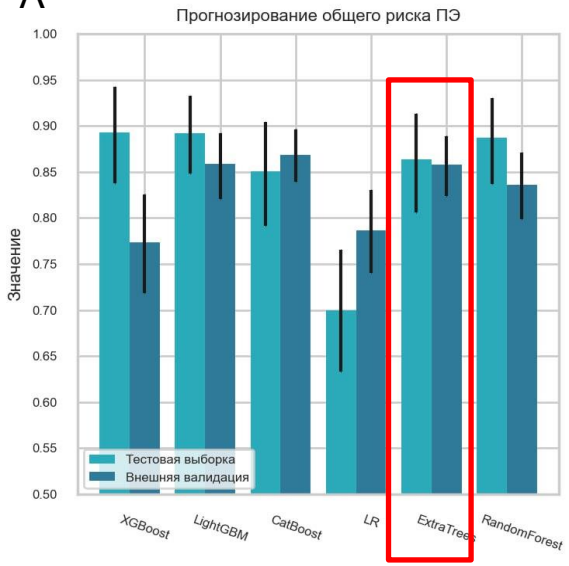
Характеристики и применение модели

- **Входные данные**: 36 признаков, включая антропометрические/лабораторные/инструментальные/анамнестические данные, симптомы
- **Выходные данные**: вероятность возникновения ранней преэклампсии, 0-100%
- **Значимые признаки:** систолическое и диастолическое АД, вес, ИМТ во время беременности, наличие отеков, задержка роста плода, возраст и рост пациентки.
- **Метрики качества на наборе данных для внешней валидации:**
  - Специфичность: **64% (62-66%) CI 95%**
  - Прогностическая ценность положительного результата (ПЦПР): **3% (2 – 4%) CI 95%**
  - Прогностическая ценность отрицательного результата (ПЦОР): **99% (99-100%) CI 95%**
- **Назначение и условия применения:**
  - применима для пациентов в возрасте от 10 до 60 лет на 7-16 неделе беременности
  - обращает внимание врача при значимых прогностических факторах ПЭ
  - предназначена для поддержки принятия врачебных решений по тактике ведения беременных

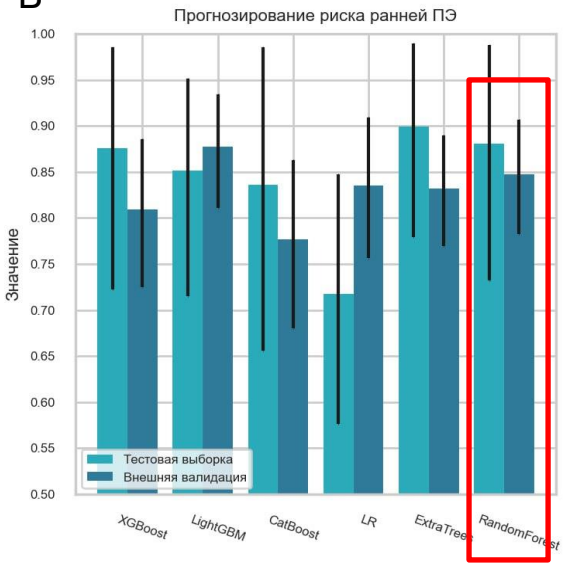
Риск преэклампсии



А



Б



Выводы

- Автоматизация процессов применима для создания предиктивных моделей для разных клинико-практических задач
- При этом учитывается не только вариатив алгоритмов машинного обучения, но и способов предобработки наборов данных, что актуально для медицинских данных реальной клинической практики
- Разработанные таким образом модели демонстрируют относительно высокую стабильность по отношению к новым данным
- Таким образом требуемые ресурсы перераспределяются от специалистов по данным к доменной экспертизе
- В дальнейшем необходимо уделять больше внимание метрикам и их совокупностям для оценки моделей и наборов данных, а так же мета-валидации моделей

Спасибо за внимание!

Есть вопросы? Пожалуйста, обращайтесь по контактам ниже:

Мои контакты

Анна Андрейченко

email: aandreychenko@webiomed.ru
telegram: @anna_medAI

Контакты компании

Сайт
https://webiomed.ru

Мы в социальных сетях

ВКонтакте

https://vk.com/webiomed

Telegram

https://t.me/webiomed

YouTube
https://www.youtube.com/@webiomed