УДК 314.48

ПРОГНОЗИРОВАНИЕ РАЗВИТИЯ ИНФАРКТА МИОКАРДА НА ОСНОВАНИИ СЕЗОННЫХ И МЕТЕОРОЛОГИЧЕСКИХ ФАКТОРОВ

© 2021 М. А. Фирюлина, И. Л. Каширина

Воронежский государственный университет (Воронеж, Россия)

В статье анализируется влияние метеорологических и сезонных факторов на частоту развития острого инфаркта миокарда (ИМ). На основании выявленных закономерностей разработаны
модели и алгоритмы прогнозирования развития ИМ на определённую дату в зависимости от сезонных характеристик и с учётом мониторинга метеорологических данных для планирования профилактических мероприятий. Анализ проводился основе данных Воронежского областного регистра ИМ за 2015-2017 годы. Для построения регрессионной модели использовались модели случайный лес и градиентный бустинг (CatBoostRegressor, LGBMRegressor, LGBMRegressor).

Ключевые слова: прогнозирование смертности, многофакторный анализ, линейная регрессия, машинное обучение.

Введение

Инфаркт миокарда (ИМ) является наиболее распространенным заболеванием среди сердечно-сосудистых заболеваний (ССЗ) [1]. Смертность от ССЗ в западных странах резко снизилась за последние десятилетия благодаря повышенному вниманию первичной профилактике, улучшению методов диагностики и лечения подобного рода заболеваний.

Инфаркт миокарда имеет сезонный характер во многих популяциях. В ряде международных исследований отмечается влияние метеорологических и сезонных факторов на развитие инфаркта миокарда [2-7]. Выявление таковой зависимости, внедрение новых технологий, позволяющих учитывать в процессе лечения все факторы, влияющие на здоровье человека, оценка риска и ущерба от климатических изменений - важные задачи, которые стоят перед современной медициной. В большинстве исследований сообщается о «зимних пиках» госпитализаций и смертности в связи с ИМ, частота событий зимой обычно на 10-20 % выше, чем во время летнего периода [2-4]. Но имеются и противоположные выводы о повышение риска смертности от ИМ в теплое время года [5]. Сезонность сердечно-сосудистых заболеваний наиболее выражена у людей, живущих в более мягком климате, которые наименее подготовлены к экстремальным погодным изменениям, поэтому часть исследований доказывает, что повышение случаев развития ИМ происходит в экстремально низкие или высокие температурные выбросы [6].

Хотя сезонные колебания сердечных заболеваний в значительной степени обусловлены предсказуемыми изменениями погодных условий, очевидна сложная взаимосвязь между условиями окружающей среды и человеком. Поведенческие и физиологические реакции на сезонные изменения модулируют восприимчивость к сезонным изменениям сердечно-сосудистой системы. Результат исследования зависит от географической расположенности региона. Люди в урбанизированной местности менее восприимчивы к перепадам погоды, чем население сельской местности [7]. Неоднородность условий окружающей среды и динамики населения во всем мире означает, что окончательное изучение этого сложного явления маловероятно.

Целью данного исследования является на основании выявленных закономерностей разработать модели и алгоритмы прогнозирования развития ИМ на определённую дату в зависимости от сезонных характеристик и с учётом мониторинга метеорологических данных для планирования профилактических мероприятий. Прогнозирование приблизи-

Фирюлина М. А. – Воронежский государственный университет, Мария Андреевна, mashafiryulina@mail.ru.
Каширина Ирина Леонидовна – Воронежский государственный университет, доктор техн. наук, kash.irina@mail.ru.

тельного числа случаев развития ИМ в определенный день поможет подготовить медицинские учреждения в случае «пиковых» ситуаций, а также предупредить пациентов, находящихся в группе риска, о более тщательном соблюдении профилактических мероприятий.

В данной статье приводятся результаты статистического анализа показателей смертности по Воронежской области, полученные путем обработки деперсонифицированных данных областного регистра инфарктов миокарда за 2015-2017 гг. Полученные результаты используются для выявления наиболее значимых факторов и построения моделей машинного обучения для прогнозирования числа случаев ИМ в определенный день. Предобработка данных проводилась с помощью СУБД Oracle 19c в среде разработки SQL Developer. Построение моделей машинного обучения, статистический анализ данных, построение графических материалов проводилось с помощью различных библиотек языка Python на платформе Goole Colab. Для построения регрессионной модели использовались встроенные пакеты языка Python для методов случайный лес и градиентный бустинг (CatBoostRegressor, LGBMRegressor, LGBMRegressor).

Статистический анализ данных

Для анализа использовалась выборка пациентов, поступивших за 2015-2017 года в больницы Воронежской области с диагнозом ИМ. Информация с неперсонифицированными данными пациентов была получена из областного регистра ИМ. Всего в исследовании было рассмотрено 11326 случая инфаркта миокарда. Средний возраст больных составил 66,5 лет. Среднее число инфарктов в день в Воронежской области (по госпитализированным пациентам) составляет 10.3433. Наибольшее число ИМ было зафиксировано в 2016 году (рис. 1).

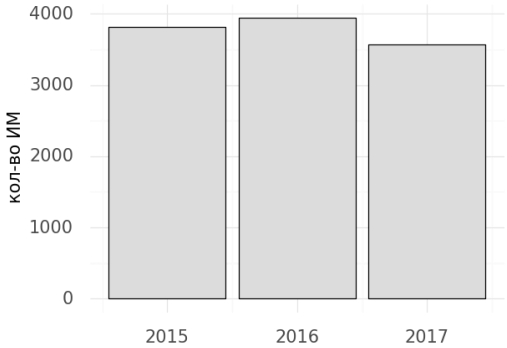

Рисунок 1. Распределение случаев ИМ по годам.

Во многих литературных источниках отмечается существенное влияние сезонности на развитие инфаркта миокарда и смертность от него. В статье [4] отмечается, что сезонность связана со значительными изменениями артериального давления. На первом этапе исследования был проведен анализ зависимостей развития ИМ по сезонам. На рисунке 2 представлена график «ящик с усами», отражающий число ИМ в Воронежской области по всем четырем временам года. “Усы” графика (I) показывают доверительный интервал, построенный с доверительной вероятностью 0.95 для среднего числа ИМ в этот сезон.

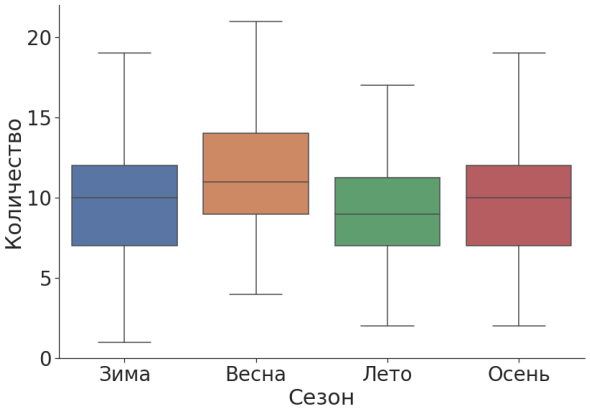

Рисунок 2. Зависимость количества ИМ от сезона.

По данному графику можно сделать вывод, что наибольшее количество инфарктов произошло весной и осенью, а не зимой или летом, когда могли наблюдаться аномально низкие или высокие температуры воздуха. Подобные результаты были выявлены в исследовании ранее [8].

Для более детального анализа, построен аналогичный график по месяцам (рис. 3).

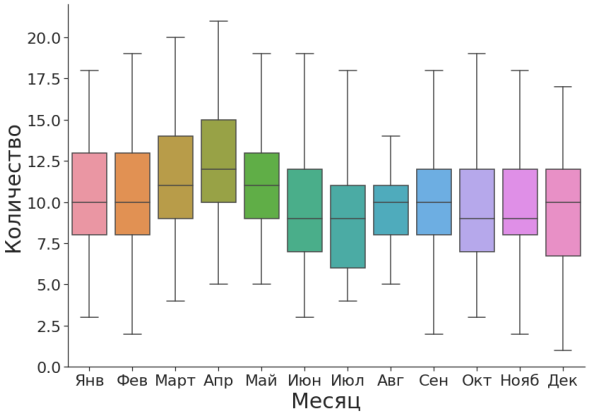

Рисунок 3. Зависимость количества ИМ от месяца.

Наибольшее количество ИМ наблюдается в апреле, а наименьшее в июле.

Чтобы понять, какие факторы имеют влияние на количество ИМ в сутки был проведен дисперсионный анализ ANOVA с помощью пакета statsmodels.formula.api. Проверялось влияние всех входных предикторов: месяц, сезон, день недели, температура воздуха, скачки температуры, атмосферное давление, скорость ветра, облачность и влажность. Тест ANOVA используется для сравнения средних значений более чем двух групп (t-критерий может использоваться для сравнения 2 групп) [9].

Таблица 1

Результаты теста ANOVA

Предиктор	P-value
MONTH	0.00001
DAY	0.024633
SEASON	0.005471
MAX_T	0. 06855
DELTA_T	0. 39385
WIND	0.74044
PRESSURE	0.622827
HUMIDITY	0.942394
CLOUDINESS	0.301945

Данный метод использует F-критерий на основе дисперсии для проверки равенства групповых средних. Иногда тест ANOVA F также называют комплексным тестом, поскольку он проверяет неспецифическую нулевую гипотезу, согласно которой все средние групп равны. По результатам, приведенным в таблице 1, видно, что наибольшее влияние оказывают сезон и месяц, и день недели.

Для уточнения был проведен анализ с помощью Т-теста. Т-тест – это тип статистики, используемый для определения значительного различия между средними значениями нескольких групп, которые могут быть связаны по определенным характеристикам [10]. В таблице 2 приведены результаты теста по месяцам. Наиболее отличительные месяцы – весенние и летние. В весеннее месяцы наблюдается повышенное количество ИМ, в летние наоборот спад.

Самые неблагоприятные месяцы – апрель и июль. Из этого можно сделать предположение, что всплеск инфарктов миокарда может быть связан с сильными перепадами температуры, которые возникают в периоды межсезонья, либо с повышением атмосферного давления.

Таблица 2

Результаты Т-теста

Месяц	Mean	Std.Err.	t-value	p-value
Янв	10.311	0.3751	-0.083	0.9333
Фев	10.647	0.4373	0.694	0.4892
Март	11.526	0.3811	3.105	0.0025
Апр	12.444	0.4160	5.050	0.0000
Май	11.279	0.3654	2.561	0.0120
Июн	9.644	0.3743	-1.866	0.0652
Июл	9.268	0.3526	-3.046	0.0030
Авг	9.419	0.3308	-2.792	0.0063
Сен	9.966	0.3842	-0.980	0.3297
Окт	10.032	0.3884	-0.800	0.4253
Нояб	10.000	0.3755	-0.914	0.3631
Дек	9.619	0.5012	-1.443	0.1522

На рисунке 4 представлен график, связывающий скачки температуры (разницу среднедневной температуры между текущим и предыдущим днем) и среднее количество инфарктов, зафиксированных в дни с таким перепадом температур. На рисунке 5 аналогичный анализ проведен для атмосферного давления. Оба предиктора были разбиты на категории для наглядности.

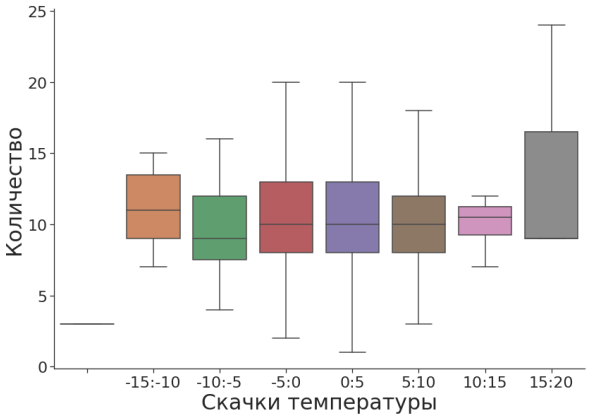

Рисунок 4. Зависимость количества ИМ от скачков температуры.

Видно, что при больших по модулю скачках, размах колебаний числа инфарктов заметно увеличивается, и максимальное число инфарктов было зафиксировано в дни, когда были достаточно сильные перепады температуры. Однако при этом связи между числом инфарктов и величиной скачка, очевидно, не наблюдается. Аномальных «пиков» относительно влияния атмосферного давления на количество ИМ не обнаружено.

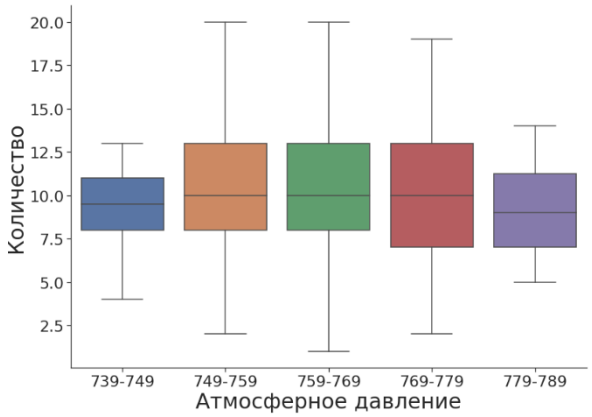

Рисунок 5. Зависимость количества ИМ от атмосферного давления.

Следующая по значимости переменная, после сезона и месяца – день недели. На рис. 6 представлен график распределения количества ИМ по дням недели. Пиковые значения приходятся на вторник, а спад наблюдается в выходные дни.

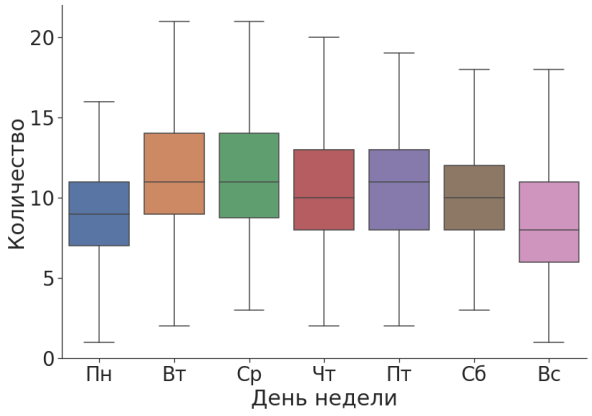

Рисунок 6. Количество ИМ по дням недели.

Построение прогнозирующей модели

Построение моделей машинного обучения и анализ данных проводились с помощью библиотек языка программирования Python. Решалась задача по прогнозированию количества ИМ на определенную дату с учетом сезонных и метеорологических факторов. На выходе прогнозировалась переменная «количество ИМ». Входные предикторы: месяц, сезон, день недели, температура воздуха, скачки температуры, атмосферное давление, скорость ветра, облачность и влажность. Для решения задачи сравнивалась точность четырех методов машинного обучения: метод случайного леса и трех методов градиентного бустинга (XGBoost, CatBoost, LGBM).

Случайный лес – алгоритм машинного обучения, заключающийся в использовании ансамбля решающих деревьев [11]. Данный алгоритм использует механизм бутстрепа, позволяющий на основе исходного обучающего набора данных с использованием случайного отбора с повторениями сформировать несколько выборок такого же размера. При построении регрессионной модели окончательная прогнозируемая величина является средним значением среди всех выходов построенных деревьев. Основным преимуществом данного алгоритма является высокая точность [12].

Градиентный бустинг – это техника машинного обучения, основная идея которой заключается в итеративном процессе последовательного построения частных моделей. Каждая новая модель обучается с использованием информации об ошибках, сделанных на предыдущем этапе, а результирующая функция представляет собой линейную комбинацию всего ансамбля моделей с учетом минимизации некоторой штрафной функции [13]. В данном исследовании были построены три модели градиентного бустинга: XGBoost, LightGBM и CatBoost. Основное различие в том, что в моделях LightGBM и CatBoost используется новая техника односторонней выборки на основе градиента (GOSS) для фильтрации экземпляров данных для нахождения разделения признаков, в то время как XGBoost использует предварительно отсортированный алгоритм и алгоритм на основе гистограмм для вычисления наилучшего разделения [14].

Для оценки точности моделей использовались метрики оценки регрессионных моделей – коэффициент корреляции, RMSE, MAE. Коэффициент корреляции показывает связь между предсказанным и реальным значением. RMSE – величина евклидова расстояния между двумя точками, прогнозируемой и исходной. Данный показатель интерпретируется как средняя ошибка модели, на сколько единиц реальное значение в среднем отличается от прогнозируемого. MAE – это линейная оценка, которая означает, что все индивидуальные различия взвешены одинаково в среднем. Преимущество данной метрики в том, что она не так чувствительна к выбросам, как RMSE.

Исходный набор данных при построении всех моделей разбивался на тестовую и обучающую выборки в соотношении 20:80. В таблице 3 приведены результаты точности полученных моделей по тестовой выборке.

Таблица 3

Метрики качества моделей

Модель	Corr	RMSE	MAE
RandomForest	0.2643	3.9437	3.0488
XGB	0.2327	4.0004	3.0974
LGBM	0.1868	4.3276	3.4462
CatBoost	0.3311	3.8644	2.992
CatBoost с лагом в 5 дней	0.4	3.6	2.8

Наилучшие показатели метрик точности продемонстрировала модель градиентного бустинга CatBoost. Для данной модели на рисунке 8 представлены наиболее значимые признаки. Можно сделать вывод, что сезонные факторы все имеют влияние, а среди метеорологических показателей наиболее значимы атмосферное давление и показания температуры воздуха.

Рисунок 8. Значимость признаков для модели CatBoost.

Во многих случаях воздействие одних факторов на другие осуществляется не мгновенно, а с некоторым временным запаздыванием – лагом. В некоторых исследованиях [3, 8] с помощью метода распределенных лагов доказано, что имеется сильная зависимость между переменными с лагом 5 дней. Для улучшения точности конечных результатов была построена модель градиентного бустинга CatBoost по входным предикторам сезонных данных и показателей атмосферного давления и температурных условий за предыдущие 5 дней. Показатель *MAE* построенной модели составляет 2.8, что говорит об уменьшении средних абсолютных разностей между целевыми значениями и прогнозами. Коэффициент корреляции составил 0.4, что говорит о том, что связь между реальным и спрогнозированным значением по тестовой выборке довольно слабая, но все-таки имеется. На рисунке 9 представлен график предсказанных и реальных значений на тестовой выборке. По графику можно отметить, что тенденция в целом повторяется, построенная модель, как правило, правильно предсказывает рост и падение количества инфарктов, но абсолютные значения она предсказывает плохо.

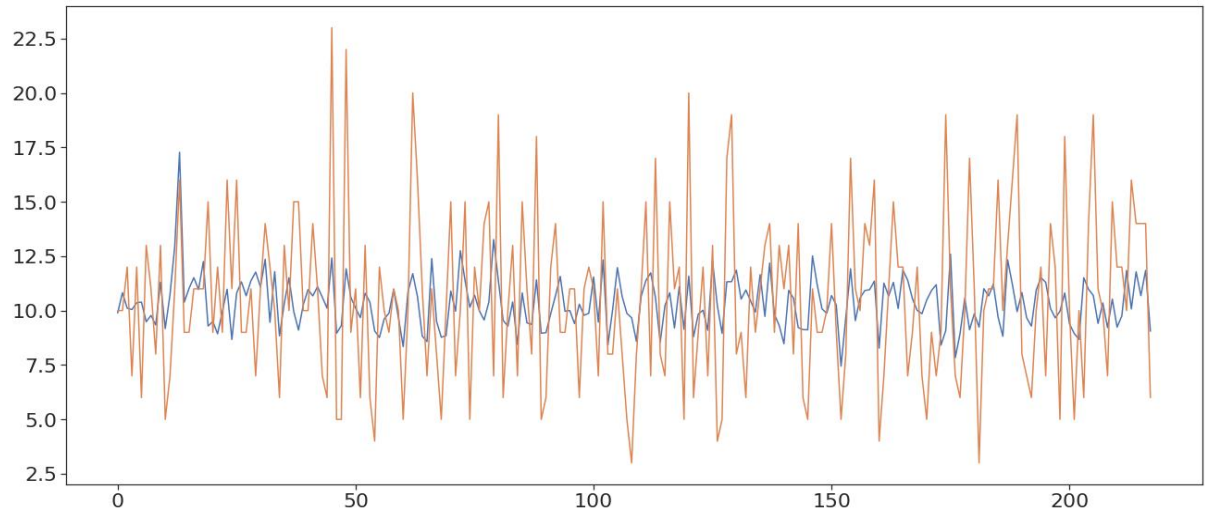

Рисунок 9. График реальных и предсказанных значений.

Таким образом, подтвердилось, что месяц, день недели, показатели температуры и атмосферное давление с некоторым лагом имеют влияние на возникновение инфаркта миокарда у населения Воронежа и Воронежской области ИМ. Однако для точного предсказания числа инфарктов в день этих данных недостаточно.

ЛИТЕРАТУРА

1. Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region // Geneva, Switzerland: World Health Organization. – 2016. – P. 201.

2. Stewart S. Seasonal variations in cardiovascular disease/ S. Srewart, A. Keates, A. Redfern // Nat Rev Cardiol. – 2017. – № 14. – P. 654-664.

3. Schneider A. Air Temperature and Inflammatory Responses in Myocardial Infarction Survivors / A. Schneider, D. Panagiotakos, S. Picciotto // Epidemiology. – 2008. – № 19 (3). – P. 391-400.

4. Cuiqing Liu Cardiovascular response to thermoregulatory challenges // Cuiqing Liu, Zubin Yavar, Qinghua Sun // Am J Physiol Heart Circ Physiol. – 2015. – № 309 (11). – P. 1793-1812.

5. Jean-Philippe Empana Increase in out-of-hospital cardiac arrest attended by the medical mobile intensive care units, but not myocardial infarction, during the 2003 heat wave in Paris, France // Jean-Philippe Empana, Patrick Sauval, Pierre Ducimetiere // Critical Care Medicine. – 2009. – № 37 (12). – P. 3079-3084.

6. Gasparrini A. Mortality risk attributable to high and low ambient temperature: a multicountry observational study / Antonio Gasparrini, Yuming Guo, Masahiro Hashizume, Eric Lavigne // The Lancet. – 2015. – № 386. – P. 369–375.

7. Urban A. Spatial Patterns of Heat-Related Cardiovascular Mortality in the Czech Republic / A. Urban, K. Burkart, J. Kyselý, C. Schuster // Int. J. Environ. Res. Public Health. – 2016. – № 13(3). – P. 284.

8. Каширина И.Л. Статистический анализ влияния метеорологических и сезонных факторов на развитие инфаркта миокарда и смертность от него по данным воронежского областного регистра / И. Л. Каширина, Р. А. Хохлов, А.О. Казакова // Врач-аспирант. – 2017. – № 85. – P. 142-150.

9. ANOVA using Python. – URL: https://www.reneshbedre.com/blog/anova.html (дата обращения: 03.06.2021).

10. Кобзарь А. И. Прикладная математическая статистика / А. И. Кобзарь. – М.: Физматлит, 2006 – 403 с.

11. Breiman L. Random Forests. Machine Learning / Leo Breiman // Machine Learning. – 2001. – № 45. – P. 5-32.

12. The Ultimate Guide to Random Forest Regression. – URL: https://www.keboola.com/blog/random-forest-regression/ (дата обращения: 10.05.2021).

13. Шитиков В.К. Классификация, регрессия, алгоритмы Data Mining с использованием R. / В.К. Шитиков, С.Э. Мастицкий – URL: https://github.com/ranalytics/data-mining/ (дата обращения: 09.05.2021).

14. Firyulina M. A. Classification of cardiac arrhythmia using machine learning techniques / M. A. Firyulina, I. L. Kashirina // J. Phys.: Conf. Ser. – 2020. – № 1614. – P. 1167-1175.

PREDICTION OF THE DEVELOPMENT OF MYOCARDIAL INFARCTION BASED ON SEASONAL AND METEOROLOGICAL FACTORS

© 2021 M. A. Firyulina, I. L. Kashirina

Voronezh State University (Voronezh, Russia)

The article analyzes the influence of meteorological and seasonal factors on the incidence of acute myocardial infarction (MI). Based on the revealed patterns, models, and algorithms for predicting the development of MI for a certain date have been developed, depending on seasonal characteristics and taking into account the monitoring of meteorological data for planning preventive measures. The analysis was carried out based on data from the Voronezh regional register of MI for 2015-2017. To build the regression model, we used random forest models and gradient boosting (CatBoostRegressor, LGBMRegressor, LGBMRegressor).

Keywords: predicting mortality, multivariate analysis, linear regression, machine learning.