УДК 004.8:616.89

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ УРОВНЯ ТРЕВОЖНОСТИ ПО ПОВЕДЕНЧЕСКИМ И СОЦИАЛЬНО-ДЕМОГРАФИЧЕСКИМ ПРИЗНАКАМ

Н.И. Липницкая
Старший преподаватель кафедры экономической информатики БГУИР
n.i.karpovich@gmail.com

М. А. Миткевич
Студент инженерно-экономического факультета БГУИР
mitkevich@bsuir.by

В. В. Смертьев
Студент инженерно-экономического факультета БГУИР
smertsyeu@bsuir.by

Н.И. Липницкая

Старший преподаватель кафедры экономической информатики УО «БГУИР». Область научных интересов: прикладные системы обработки данных, информационные технологии и программирование, методика преподавания.

М. А. Миткевич

Студент инженерно-экономического факультета УО «БГУИР». Область научных интересов: машинное обучение, анализ данных, поведенческая аналитика.

В. В. Смертьев

Студент инженерно-экономического факультета УО «БГУИР». Область научных интересов: анализ данных, поведенческая аналитика, методы машинного обучения.

Аннотация. Рост распространённости тревожных расстройств требует разработки доступных инструментов скрининга. В данном исследовании проверяется гипотеза о возможности косвенной оценки уровня тревожности по GAD-7 на основе поведенческих и социально-демографических факторов. На основе данных опроса 439 респондентов обучены и протестированы модели машинного обучения (Ridge, Random Forest, CatBoost, LightGBM). CatBoost продемонстрировал наилучшие результаты: MAE = 2.79 балла, R^2 = 0.39. Ключевыми протективными предикторами являются удовлетворённость рабочей/учебной средой и качество сна, факторами риска – дефицит социальной поддержки и негативные жизненные события.

Ключевые слова: тревожность, психическое здоровье, оценка риска, машинное обучение, поведенческие факторы, цифровая психиатрия, медицинская информатика, GAD-7, поведенческая аналитика.

Введение. Цифровая трансформация здравоохранения (e-Health) смещает акцент с лечения уже развившихся заболеваний на профилактику и ранний скрининг. Психические расстройства, в том числе генерализованное тревожное расстройство, остаются одной из ключевых причин снижения качества жизни, при этом тревожные состояния затрагивают значительную часть трудоспособного населения. Их диагностика нередко запаздывает из-за стигматизации и нехватки специалистов.

В настоящее время основным инструментом скрининга тревожных расстройств являются стандартизированные опросники, такие как GAD-7 [2; 3]. Несмотря на высокую надёжность, они предполагают прямой контакт с врачом или осознанное и регулярное заполнение пациентом, что ограничивает их применение в массовых профилактических программах. Развитие методов машинного обучения открывает возможность перехода к «цифровому фенотипированию», когда уровень риска оценивается по косвенным, но легко собираемым данным.

Модели машинного обучения уже применяются в психологии и психиатрии для прогнозирования депрессии, тревожности и стресса по разнородной информации: текстам, активности в социальных сетях, физиологическим сигналам и цифровым логам [4; 5]. Однако такие подходы часто требуют специального оборудования, длительного мониторинга или вызывают серьёзные вопросы конфиденциальности. Большая часть работ опирается либо на дорогостоящие биомедицинские сигналы (например, ЭЭГ, вариабельность сердечного ритма) , либо на пассивный цифровой след [7; 8], в то время как потенциал моделей, использующих только целенаправленно собранные поведенческие факторы, пока изучен недостаточно.

В данной работе проверяется гипотеза о том, что совокупность простых, легко собираемых поведенческих паттернов и социально-демографических характеристик может обеспечить приемлемую точность прогноза уровня тревожности для задач первичного скрининга. Цель исследования – разработать и оценить модель машинного обучения на основе алгоритмов Ridge, Random Forest, CatBoost и LightGBM и сравнить их качество. В отличие от многих исследований, использующих дорогие медицинские измерения или пассивные цифровые следы, здесь прогноз строится только по анкетным данным, не требующим сложной инфраструктуры.

Дополнительным вкладом работы является подробный анализ ошибок модели в разных диапазонах тревожности и интерпретация значимости признаков с помощью метода SHAP. В качестве ориентира для оценки приемлемой точности используются минимально клинически значимая разница по шкале GAD-7 (3–4 балла) и типичный диапазон значений коэффициента детерминации R^2 для психометрических моделей на опросных данных.

Организация исследования и сбор данных. Сбор данных проводился в формате одномоментного онлайн-опроса с использованием платформы Google Forms. В исследовании приняли участие 439 респондентов, преимущественно студенты и молодые специалисты. Форма была настроена так, чтобы исключить пропуски ответов, поэтому в

итоговом датасете отсутствовали пропущенные значения. В качестве целевой переменной использовалась суммарная оценка по шкале GAD-7 (от 0 до 21 балла). Предикторы были получены на основе авторских вопросов, опирающихся на психометрические исследования образа жизни, и включали показатели качества сна, уровня физической активности, особенностей питания, наличия вредных привычек, социальной поддержки и удовлетворённости учебной/рабочей средой.

Описание признакового набора. Предварительная обработка данных и построение моделей выполнялись на языке Python с использованием библиотек Pandas и Scikit-learn. Все переменные были переведены в числовой формат, пропусков в данных не осталось. Категориальные и порядковые признаки были закодированы так, чтобы сохранять естественный порядок их значений. Исходные предикторы объединены в три смысловые группы, показанные в таблице 1.

Таблица 1. Категории факторов и их описание

Категория факторов	Описание предикторов
Физиологические и поведенческие	Продолжительность сна, уровень физической активности, качество питания, статус курения, употребление алкоголя, приём лекарственных препаратов.
Социально-психологические	Удовлетворённость учебной или рабочей средой, субъективный уровень социальной поддержки, посещение психолога, наличие травмирующих событий за последний год.
Рутинные паттерны	Экранное время (использование гаджетов), частота прогулок на свежем воздухе, хронотип (суточный ритм активности).

Анализ взаимосвязи данных. На этапе предварительного исследования данных был проведён корреляционный анализ для выявления линейных зависимостей между предикторами и целевой переменной.

Разработка предиктивных моделей машинного обучения. Разработка и обучение моделей проводились в среде Google Colab на языке Python с использованием библиотек Scikit-learn, LightGBM и CatBoost. Для решения задачи были выбраны алгоритмы разной природы: ансамблевые методы Random Forest, LightGBM и CatBoost, а также линейная Ridge-регрессия, использовавшаяся как базовая модель для сравнения.

Отдельно тестировалась стратегия логарифмического преобразования целевой переменной (Log-Target), однако во всех случаях это приводило к ухудшению качества (росту ошибок), поэтому данный подход был отклонён.

Оценка качества моделей выполнялась на отложенной тестовой выборке (hold-out) в пропорции 77,5% / 22,5% с использованием стратифицированного разбиения по уровням тревожности, чтобы сохранить структуру распределения GAD-7. Основной метрикой выступала средняя абсолютная ошибка (MAE), измеряемая в баллах шкалы GAD-7. Дополнительно рассчитывались коэффициент детерминации R^2 и среднеквадратическая ошибка (RMSE) для более полной оценки качества прогноза.

Результаты эксперимента и интерпретация итоговой модели. По результатам серии экспериментов наивысшую предсказательную точность продемонстрировала модель CatBoost с оптимизацией гиперпараметров методом Optuna. Математически предсказание модели CatBoost представляет собой взвешенную сумму предсказаний T деревьев решений:

$$YGAD7=t=1ThtX \quad (1)$$

где YGAD7 – прогнозируемый уровень тревожности, htX – предсказание t-го дерева на объекте X, η – скорость обучения (learning rate), T – число деревьев в ансамбле.

Сводные результаты тестирования лучших конфигураций моделей представлены в таблице 2.

Таблица 2. Сводные результаты тестирования лучших конфигураций моделей

Алгоритм	Конфигурация	MAE	RMSE	R^2
CatBoost	Optuna	2.79	3.65	0.39
Ridge Regression	GridSearch (α=10)	2.87	3.71	0.36
LightGBM	Optuna	2.87	3.70	0.37
Random Forest	Optuna + Gap Penalty	2.87	3.71	0.36

По результатам сравнительного эксперимента наилучшие показатели продемонстрировала модель CatBoost с подбором гиперпараметров (Optuna): MAE = 2.79, R^2 = 0.39. Линейная Ridge-регрессия показала очень близкие результаты (MAE = 2.87, R^2 = 0.36), что косвенно указывает на преобладание аддитивных эффектов предикторов и относительно небольшую роль сложных нелинейных взаимодействий. Логарифмическое преобразование целевой переменной и метод ансамблирования на основе случайных подвыборок не дали прироста качества и не были использованы в финальной конфигурации моделей.

Интерпретация предикторов. Для содержательной интерпретации модели CatBoost применялся метод SHAP (SHapley Additive exPlanations), позволяющий оценить вклад каждого признака в индивидуальное предсказание (рис. 1). В отличие от весовых коэффициентов линейных моделей, SHAP-значения корректно отражают нелинейные зависимости и взаимодействия между признаками.

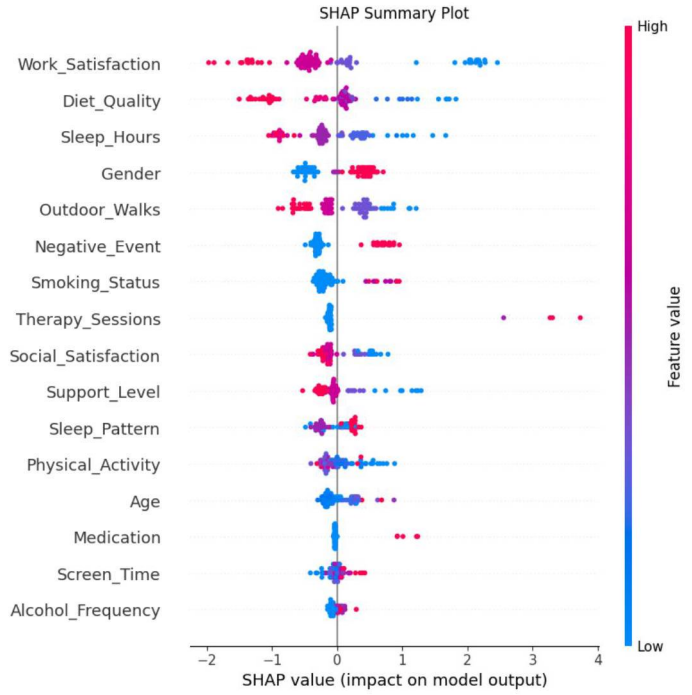
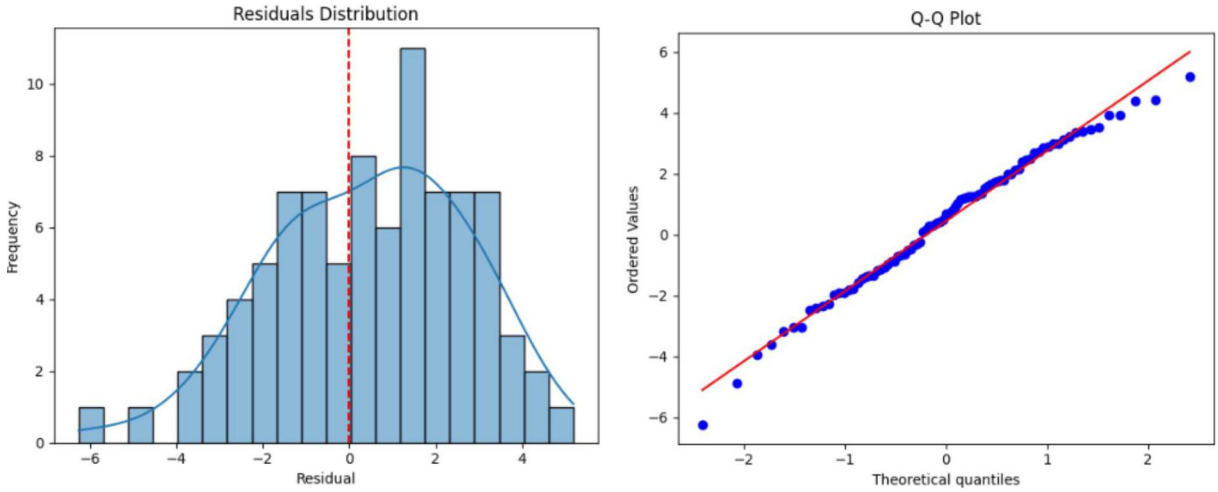


Рисунок 1. SHAP Summary Plot модели CatBoost: каждая точка – один респондент

Анализ SHAP Summary Plot показывает, что наибольший защитный эффект имеет удовлетворённость текущей деятельностью (Work_Satisfaction): высокие значения этого признака смещают предсказание в сторону более низкого уровня тревожности. Аналогичный протективный характер демонстрируют качество сна (Sleep_Hours) и частота прогулок (Outdoor_Walks). Фактором риска выступает посещение психолога (Therapy_Sessions), однако эту связь можно рассматривать как проявление обратной причинности: люди с более высокой тревожностью чаще обращаются за профессиональной помощью.

Анализ ошибок и диагностика остатков. Для оценки надёжности модели был проведён детальный анализ остатков. Гистограмма распределения остатков приближена к нормальному закону, что свидетельствует об отсутствии сильного систематического смещения. Анализ ошибок по диапазонам тревожности выявил ограничения: для уровней Low (0–5 баллов) и Mid-Low (5–10 баллов) модель работает с высокой точностью (MAE 1.9–2.9 балла), тогда как для уровня High (>15 баллов) ошибка резко возрастает до MAE ≈ 6.2 балла. Модель склонна недооценивать экстремально высокие уровни тревожности, сглаживая их к средним показателям выборки.



а *б*

Рисунок 2. Диагностика остатков модели CatBoost: а – гистограмма распределения остатков; б – Q-Q plot

Характер систематических отклонений наглядно иллюстрирует рис. 3, на котором представлено сравнение предсказанных и истинных значений GAD-7 на тестовой выборке.

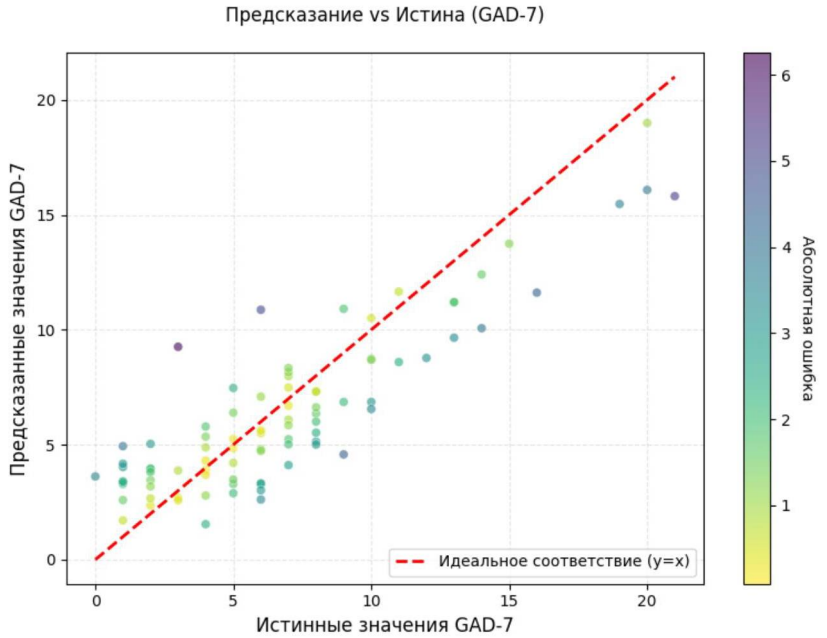


Рисунок 3. Сравнение предсказанных и истинных значений GAD-7 на тестовой выборке

Анализ рис. 3 показывает два устойчивых типа ошибок. В диапазоне низких значений тревожности (0–4 балла) модель чаще завышает прогноз: точки располагаются выше линии идеального соответствия, что связано с эффектом регрессии к среднему при малом числе таких наблюдений. В области высоких значений (>15 баллов), напротив, модель систематически занижает тревожность, что видно по смещению точек ниже диагонали. Тёмно-синие точки с абсолютной ошибкой 5-6 баллов соответствуют ранее полученному MAE ≈ 6,2 для диапазона High. Это ограничение связано с недостаточным числом респондентов с высокой тревожностью в обучающей выборке и указывает на необходимость целевого расширения датасета в этом сегменте.

Интерпретация результатов. Полученное значение MAE = 2,79 балла важно не только с технической, но и с клинической точки зрения. В психометрии минимально клинически значимая разница (MCID) для шкалы GAD-7 составляет примерно 3-4 балла. То, что средняя абсолютная ошибка модели не превышает этот диапазон, указывает на её потенциальную пригодность как вспомогательного инструмента для мониторинга изменений уровня тревожности во времени, особенно в исследовательских и пилотных проектах.

Коэффициент детерминации $R^2 = 0,39$ также можно считать достаточно высоким для поведенческих наук, где при использовании только опросных данных значения R^2 обычно не выходят за пределы 0,30-0,40. При этом остаётся важное ограничение: точность прогноза заметно снижается для высоких значений шкалы (более 15 баллов), что связано с малым числом респондентов с выраженной тревожностью в выборке. Для компенсации этого дисбаланса была протестирована техника синтетического оверсэмплинга для регрессии (SMOGN), однако она не дала улучшения глобальных метрик качества на тестовой выборке. Дальнейшие исследования целесообразно ориентировать на целевое расширение выборки в сегменте высоких значений GAD-7 и более детальную оценку точности моделей в отдельных подгруппах по уровню тревожности. Это позволит точнее оценить применимость разработанного подхода для задач раннего выявления групп риска.

Заключение. В ходе исследования была разработана и протестирована модель машинного обучения для оценки уровня тревожности на основе поведенческих и социально-демографических факторов. Сравнительный анализ показал, что на выборке ограниченного объёма наилучшие результаты демонстрирует модель CatBoost с оптимизацией гиперпараметров (MAE = 2.79, R^2 = 0.39), тогда как линейная Ridge-регрессия обеспечивает очень близкую точность при более простой интерпретации.

Полученные результаты показывают, что на основе относительно простых опросных данных возможно построить модель, пригодную для вспомогательной оценки уровня тревожности и предварительного выделения групп повышенного риска, в том числе в формате цифровых инструментов, интегрируемых в образовательные порталы или мобильные приложения. В то же время снижение точности в диапазоне высоких значений GAD-7 подчёркивает необходимость расширения выборки за счёт респондентов с выраженной тревожностью и проведения дополнительной валидации на независимых выборках. Перспективными направлениями дальнейшей работы являются: целевое пополнение данных респондентами с высоким уровнем тревожности, проверка модели в других возрастных и профессиональных группах, а также исследование комбинированных подходов, объединяющих опросные данные с другими источниками информации (например, пассивными цифровыми следами или физиологическими показателями) для повышения чувствительности и специфичности скрининга.

Список литературы

[1] Gray, B., Asrat, B., Brohan, E., Chowdhury, N., Dua, T., & van Ommeren, M. (2024). Management of generalized anxiety disorder and panic disorder in general health care settings: new WHO recommendations. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, *23*(1), 160–161. https://doi.org/10.1002/wps.21172

[2] Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

[3] Kliem, S., Sachser, C., Lohmann, A., Baier, D., Brähler, E., Fegert, J. M., & Gündel, H. (2025). Psychometric evaluation and community norms of the GAD-7, based on a representative German sample. *Frontiers in psychology*, *16*, 1526181. https://doi.org/10.3389/fpsyg.2025.1526181

[4] Мосолова Е.С., Алфимов А.Е., Костюкова Е.Г., Мосолов С.Н. Перспективы применения методов машинного обучения при аффективных расстройствах // Digital Diagnostics. 2025. Т. 6. №1. С. 97–115. doi: 10.17816/DD634885

[5] Respiration Rate and Volume Measurements Using Wearable Strain Sensors / M. Chu [et al.] // npj. Digital Medicine. 2019. No 2. https://doi.org/10.1038/s41746-019-0083-3.

[6] Устройство для измерения активной и емкостной составляющих импеданса биологических тканей: пат. 2196504 Рос. Федерации, МПК A 61 B 5/053 / А. В. Ефремов, Р. Р. Ибрагимов, Р. А. Манвелиадзе, В. Т. Леонтьев, К. Г. Булатецкий, Г. Г. Колонда, Е. В. Тарасов, Р. Ш. Ибрагимов; Новосиб. гос. мед. акад., № 2000117324/14. Заявл. 28.06.2000. Опубл. 20.01.2003.

[7] Программная модель системы для анализа импедансометрических характеристик биологических жидкостей / К. Е. Мешкова [и др.] // «Медэлектроника–2022. Средства медицинской электроники и новые медицинские технологии»: сб. науч. ст. XIII Междунар. науч.-техн. конф., г. Минск, 8–9 декабря 2022 г. Минск: Белор. гос. ун-т информ. и радиоэлек., 2022. С. 93–97.

[8] СКОРЫНИН А. А. ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ. ПЕДАГОГИКА И ПСИХОЛОГИЯ //ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ. ПЕДАГОГИКА И ПСИХОЛОГИЯ Учредители: Пермский государственный гуманитарно-педагогический университет. – №. 9. – С. 58-65.

[9] Toussaint A., et al. Sensitivity to change and minimal clinically important difference of the GAD-7. Journal of Affective Disorders, 2020, vol. 265, pp. 395-401.

[10] Oh C-M, Kim HY, Na HK, Cho KH and Chu MK (2019) The Effect of Anxiety and Depression on Sleep Quality of Individuals With High Risk for Insomnia: A Population-Based Study. Front. Neurol. 10:849. doi: 10.3389/fneur.2019.00849.

Авторский вклад

Липницкая Наталья Ивановна – постановка общей научной задачи исследования, формирование концепции работы и структуры статьи, контроль методологии построения и валидации моделей машинного обучения, научное редактирование текста, подготовка статьи к публикации и доработка выводов.

Миткевич Максим Андреевич – сбор и предобработка данных, разработка и реализация инструментария сбора данных (Google Forms), обучение и тестирование моделей машинного обучения, анализ предикторов.

Смертьев Владислав Валерьевич – разработка методологии, обучение и тестирование моделей машинного обучения, анализ предикторов, анализ ошибок и диагностика остатков модели, обсуждение результатов, написание статьи и оформление статьи

MACHINE LEARNING MODELS FOR PREDICTING ANXIETY LEVELS FROM BEHAVIORAL AND SOCIO-DEMOGRAPHIC FEATURES

N.I. Lipnitskaya
Senior Lecturer of the Department of Economic Informatics, BSUIR
n.i.karpovich@gmail.com

M. A. Mitkevich
Student of the Faculty of Engineering and Economics, BSUIR
mitkevich@bsuir.by

U. V. Smertsyeu
Student of the Faculty of Engineering and Economics, BSUIR
smertsyeu@bsuir.by

Abstract. The increasing prevalence of anxiety disorders requires the development of accessible screening tools. This study tests the hypothesis that it is possible to indirectly assess anxiety levels on the GAD-7 scale from behavioral and socio-demographic factors, with GAD-7 used solely as the target variable and not as part of the input questionnaire. Based on survey data from 439 respondents, several machine learning models (Ridge, Random Forest, CatBoost, LightGBM) were trained and tested. CatBoost demonstrated the best results: MAE = 2.79, R^2 = 0.39. Key protective predictors are satisfaction with the work/study environment and sleep quality; risk factors include lack of social support and negative life events.

Keywords: anxiety, mental health, risk assessment, machine learning, behavioral factors, digital psychiatry, medical informatics, GAD-7, behavioral analytics.