Предсказание результатов теста Р. Кеттела на основе подписок пользователя в социальной сети

Г. Е. Рязанцев[1], В. Д. Олисеенко[2], М. В. Абрамов[3]
[1]*Санкт-Петербургский государственный университет*
[2,3] *Санкт-Петербургский Федеральный исследовательский центр Российской академии наук*
[1]st088141@student.spbu.ru, [2]vdo@dscs.pro, [3]mva@dscs.pro

***Аннотация*. В данной работе рассматривается взаимосвязь между личностными особенностями пользователя, оцениваемой с помощью теста Р. Кеттела и его подписками в одной из российских социальных сетей. По полученным результатам удалось показать, что некоторые факторы возможно предсказать лучше, чем случайным образом (факторы A, B, C, F, G, N, O, Q1, Q3, Q4 теста Р. Кеттела). Результаты данного исследования могут быть использованы для разработки более сложной модели, учитывающей больше пользовательских данных.**

Ключевые слова: мультиклассовая классификация, машинное обучение, 16-факторный тест Кеттела, социальные сети, уменьшение пространства

I. Введение

Социальную сеть «ВКонтакте» ежедневно посещают миллионы пользователей (по данным https://vk.com/2022), при этом оставляя там огромное количество информации о себе: личные данные (факты из биографии), интересы и предпочтения, посты, видео, картинки, подписки на группы и т. д. Эта информация косвенно может отражать различные личностные особенности человека [1], владельца аккаунта социальной сети. Информация о личностных особенностях может быть использована во многих сферах жизни: образовательной [2], банковской [4], медицинской [5], рекомендательной [6] и др. При этом с помощью алгоритмов можно предсказывать личностные особенности большого количества пользователей, тратя на это в разы меньше времени и средств, чем при тестировании каждого человека в отдельности с использованием психологических тестов.

Цель работы заключается в выявления взаимосвязи (с использованием методов машинного обучения) между результатами 16-факторного теста Р. Кеттела и подписками на группы пользователей в социальной сети. Для достижения данной цели были собраны подписки пользователей, которые прошли тест Р. Кеттела, данные о подписках были преобразованы в формат, подходящий для анализа и использования методов машинного обучения (логистическая регрессия, случайный лес, метод опорных векторов, градиентный бустинг над решающими деревьями). Теоретическая значимость работы заключается в выработке методологии для проверки взаимосвязи результатов психологического тестирования (по 16-факторному тесту Р. Кеттела) и подписок пользователей в социальной сети. Практическая значимость заключается в разработке программной платформы, которая может лечь в основу автоматизированной системы предсказания результатов 16-факторному тесту Р. Кеттела по странице пользователя в социальной сети.

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № FFZF-2022–0003; при финансовой поддержке гранта Президента Российской Федерации МК-5237.2022.1.6.

II. Релевантные работы

В работах [7]–[10], связанных с предсказанием личностных особенностей на основе подписок пользователей можно выделить два подхода. Первый подход, рассмотренный в статьях [7], [8], заключается в представлении подписок пользователей в виде матрицы, где в столбцах расположены группы, а в строках пользователи. На пересечении строк и столбцов ставились единицы, если пользователь подписался на группу, иначе 0. Далее к этой матрице были применены различные алгоритмы машинного обучения. Второй подход, применённый в статье [10], заключается в выделении различных признаков у каждой группы, на которые подписан пользователь, а затем вычислении для каждого пользователя вектора, который описывает его подписки. Далее на полученных векторах предсказывались личностные особенности пользователей.

В статье [7] авторы рассматривали вопрос предсказания личностных особенностей пользователей, оцениваемой с помощью теста Большая пятёрка, на основе лайков, которые были поставлены группам в социальной сети «Facebook» (аналог подписок в социальной сети «ВКонтакте»). Для представления подписок авторы использовали первый подход (основанный на матрицах), после чего получали разреженную матрицу и уменьшали размерность матрицы с использованием сингулярного разложение (SVD). В качестве метода машинного обучения предсказания личностных особенностей (результатов теста Большой пятерки) пользователей была использована логистическая регрессия. Точность измерялась с помощью корреляции Пирсона между предсказанными и реальными значениями. В результате лучше всего получилось предсказывать черту личности «Открытость»: корреляция составила 0,43. Другие личностные качества показали результат от 0,29 до 0,4. В статье [8] также использовали первый подход для предсказания депрессии. В качестве классификаторов авторы использовали следующие модели: метод опорных векторов, случайный лес, логистическая регрессия, многослойный перцептрон, наивный байесовский классификатор, алгоритмы градиентного бустинга (Adaptive Boosting, XGBoost, LightGBM

Boosting). Качество измерялось с помощью F1-меры. Лучший результат показал наивный байесовский классификатор с F1-мерой, равной 0,55. Но после уменьшения размерности с помощью метода главных компонент получили точность 0,65 при использовании случайного леса.

В статье [10] автором был рассмотрен вопрос предсказания депрессии на основе цифровых следов в социальной сети «ВКонтакте». Для построения предсказательной модели на основе подписок пользователей было предложено представить каждого пользователя в виде вектора в 28-мерном пространстве. Каждая координата вектора представляла количество подписок пользователя, которые составлялись из следующих признаков: размер группы (маленькая, средняя, большая), возрастные ограничения (0+, 16+, 18+, не указано) и тематика группы (юмор, творчество, кулинария, образование, медиа, городское сообщество, шоу, литература, общество, наука, дизайн, неопределенный тип сообщества, культура, кино, стиль, фотография, туризм, музыка, художник, животные и уход за собой). Для предсказания были рассмотрены следующие методы машинного обучения: градиентный бустинг (XGBoost, Light-GBM, CatBoost), метод опорных векторов, случайный лес, наивный байесовский классификатор, метод ближайших соседей, многослойный перцептрон, логистическая регрессия. Качество измерялось с помощью метрики F1-macro. Лучший результат показал метод опорных векторов со значением 0,67. Таким образом, задача предсказания личностных особенностей пользователя на основе подписок имеет несколько подходов к решению, однако ни один из них не был применён к задаче предсказания результатов тест Р. Кеттела.

III. Постановка задачи

В данной работе рассматривается тест Р. Кеттела, который состоит из 16 факторов (A: открытость, B: уровень интеллекта, C: эмоциональная стабильность, E: доминантность, F: экспрессивность, G: моральная нормативность, H: смелость, I: эмоциональная чувствительность, L: подозрительность, M: мечтательность, N: дипломатичность, O: тревожность, Q1: восприимчивость к новому, Q2: самостоятельность, Q3: самодисциплина, Q4: напряжённость). Тест состоит из 167 вопрос, каждый вопрос содержит три ответа. Результаты тестирования представляются в виде оценки по 10 балльной шкале (от 1 до 10) для каждому из факторов. Несмотря на получаемую оценку, часто психологам может быть важен не сам конкретный балл, а общая его направленность в меньшую или большую сторону. Таким образом можно свести задачу предсказания результатов теста Р. Кеттела к 16 задачам трёх классовой (мультиклассовой) классификации, где класс определяет низкую выраженность фактора (оценка от 1 до 4), отсутствие выраженности (от 5 до 6) и высокую выраженность фактора (от 7 до 10). Математическая постановка задачи определяется следующим образом: пусть X — множество групп пользователя, $y = \{0, 1, 2\}$ — степень выраженности фактора. Тогда для каждого пользователя необходимо построить 16 таких классификаторов *Fi*, что каждый из 16 классификаторов *Fi(X)* получает на вход подписки пользователя и на выход выдаёт степень выраженности (класс) *i*-того фактора.

IV. Описание решения задачи

В данном исследовании использовались обезличенные данные пользователей, которые прошли тест Р. Кеттела в приложении[1] во «ВКонтакте». Всего было собрано 300 записей, каждая из которых представляла собой список групп пользователя и результат прохождения теста Р. Кеттела. Эксперимент выполнялся с использованием языка программирования Python 3.9.10 совместно со следующими библиотеками: Numpy 1.22.2, Pandas 1.4.0, Scikit-learn 1.1.2, XGBoost 1.7.0, CatBoost 1.1.1.

Дальнейшая работа строилась по следующему алгоритму принципу. У каждой группы были собраны некоторые данные (количество участников, возрастные ограничения, тематика и информация о том, закрытая или открытая группа). При анализе информации оказалось, что тематик у групп более 400, поэтому было принято решение уменьшить размерность, разделив темы на кластеры. Каждая тематика была представлена в виде вектора с помощью предобученной модели fastText[2]. Затем данные вектора были поданы на вход алгоритму К-средних (K = 36), в результате работы которого были темы были разбиты на 36 кластеров. После этого было удалено 23 пользователя, у которых не было подписок и для каждого пользователя был вычислен 43-мерный вектор, каждая координата вектора представляла количество групп, которые относятся к одному из следующих признаков (один из 36 классов тематик, размер группы (маленькая, средняя, большая), возрастные ограничения (0+, 16+, 18+, не указано).

Для дальнейшей классификации были рассмотрены следующие алгоритмы машинного обучения с различными гиперпараметрами: логистическая регрессия с параметрами по умолчанию, метод опорных векторов с параметрами по умолчанию, случайный лес со следующими параметрами:

- количество деревьев от 10 до 100 с шагом 15,
- максимальная глубина от 2 до 15 с шагом 2,
- минимальное количество экземпляра для конечного узла от 1 до 8 с шагом 2,
- минимальное количество экземпляров для разделения узла от 2 до 10 с шагом 3

и две реализации градиентного бустинга над решающими деревьями (XGBoost и CatBoost) со следующими параметрами:

- количество деревьев от 10 до 200 с шагом 50,
- максимальная глубина от 2 до 11 с шагом 3,
- темп обучения: 0,05, 0,3, 0,5,
- минимальное количество экземпляра для конечного узла от 1 до 8 с шагом 2.

Для подбора гиперпараметров использовался метод gridsearch из библиотеки scikit-learn, который позволяет перебрать все заданные комбинации и оценить их при помощи перекрёстной проверки на пяти блоках тестовой

[1] «Психологические тесты. URL: https://vk.com/ticspsytests (дата обращения 18.03.2023). »

[2] «fastText. URL: https://fasttext.cc (дата обращения: 18.03.2023)»

выборки. Чтобы оценить результат классификации использовались следующие метрики: F1-micro (среднее арифметическое F1-мера для каждого класса), F1-macro (F1-мера для всей выборки), F1-weighted (среднее взвешенное F1-меры для каждого класса), которые являются релевантными для задач мультиклассовой классификации [11].

$$\text{F1-micro} = \frac{2(\text{TP}_1 + \text{TP}_2 + \text{TP}_3)}{2(\text{TP}_1 + \text{TP}_2 + \text{TP}) + (\text{FP}_1 + \text{FP}_2 + \text{FP}) + (\text{FN}_1 + \text{FN}_2 + \text{FN}_3)}$$

$$\text{F1-weighted} = \frac{N_1}{N} \cdot \frac{2\text{TP}_1}{2\text{TP}_1 + \text{FP}_1 + \text{FN}} + \frac{N_2}{N} \cdot \frac{2\text{TP}_2}{2\text{TP}_2 + \text{FP}_2 + \text{FN}_2} + \frac{N_3}{N} \cdot \frac{2\text{TP}_3}{2\text{TP}_3 + \text{FP}_3 + \text{FN}_3}$$

$$\text{F1-macro} = \frac{1}{3} \cdot \left(\frac{2\text{TP}_1}{2\text{TP}_1 + \text{FP}_1 + \text{FN}} + \frac{2\text{TP}_2}{2\text{TP}_2 + \text{FP}_2 + \text{FN}_2} + \frac{2\text{TP}_3}{2\text{TP}_3 + \text{FP}_3 + \text{FN}_3} \right)$$

где TP_1, TP_2, TP_3 — истинно положительные ответ для каждого из трёх классов, FP_1, FP_2, FP_3 – ложно положительные ответ для каждого из трёх классов, FN_1, FN_2, FN_3 – ложно отрицательные ответ для каждого из трёх классов. N_1, N_2, N_3 – количество экземпляров каждого из трёх классов, N – общее количество экземпляров.

V. Эксперимент

В таблице I представлены результаты эксперимента. В каждой строке выписан фактор личности, лучшие результаты метрик и модель, на которой удалось достичь таких результатов.

ТАБЛИЦА I. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Фактор	Модель	F1-macro	F1-micro	F1-weigh
A	CatBoost	**0,351**	**0,438**	**0,425**
B	XGBoost	**0,342**	**0,483**	**0,454**
C	CatBoost	**0,384**	**0,461**	**0,429**
E	RF	0,315	0,461	0,441
F	CatBoost	**0,336**	**0,539**	**0,456**
G	XGBoost	**0,34**	**0,64**	**0,579**
H	XGBoost	0,323	0,359	0,353
I	XGBoost	0,314	0,483	0,464
L	RF	0,332	0,472	0,476
M	CatBoost	0,311	0,472	0,419
N	CatBoost	0,359	0,783	0,656
O	CatBoost	**0,48**	**0,584**	**0,569**
Q1	XGBoost	**0,36**	**0,472**	**0,454**
Q2	XGBoost	0,319	0,483	0,46
Q3	XGBoost	**0,422**	**0,652**	**0,653**
Q4	XGBoost	**0,395**	**0,573**	**0,547**

По результатам эксперимента можно сделать вывод, что лишь на некоторых личностных факторах (A, B, C, F, G, N, O, Q1, Q3, Q4) удалось добиться качества, превышающего случайную классификацию. Низкий результат классификации мог быть вызван небольшим размером обучающей выборки. При этом при обучении отслеживались все три метрики, чтобы избежать ситуации, когда классификатор отлично распознаёт одни классы, при этом, не распознавая остальные (видно по метрики F1-macro).

VI. Заключение

В работе была рассмотрена возможность предсказание результатов тест Р. Кеттела на основе подписок пользователя. Проведено обучение и сравнение различных классификаторов и на основе результатов предсказаний показано, что с помощью алгоритмов машинного обучения можно предсказывать личностные факторы пользователя, хоть и с небольшой точностью. Результаты данного исследования могут быть использованы для разработки более сложной модели, учитывающей большее количество пользовательских данных.

Теоретическая значимость работы заключается в выработке методология для проверки взаимосвязи результатов психологического тестирования (по 16 факторному тесту Р. Кеттела) и подписок пользователей в социальной сети. Практическая значимость заключается в разработке программной платформы, которая может лечь в основу автоматизированной системы предсказания результатов 16 факторному тесту Р. Кеттела по странице пользователя в социальной сети.

Список литературы

[1] Golbeck J., Robles C., Turner K. Predicting Personality with Social Media // CHI'11 Extended Abstracts on Human Factors in Computing. 2011. pp. 253–262. Doi: 10.1145/1979742.1979614

[2] De Raad B., Schouwenburg H. Personality in learning and education: a review // European Journal of Personality. 1996. № 5. pp. 303–336. Doi: 10.1002/(SICI)1099-0984(199612)10:5<303::AID-PER262>3.0.CO;2-2

[3] Krylov B., Abramov M., Khlobystova A. Automated Player Activity Analysis for a Serious Game About Social Engineering // Studies in Systems, Decision and Control. 2021. Vol. 337. P. 587–599 Doi: 10.1007/978-3-030-65283-8_48

[4] Ho S. P. S., Wong A. The role of customer personality in premium banking services // Journal of Financial Services Marketing. 2022. P. 1–21. Doi: 10.1057/s41264-022-00150-3

[5] Kern M.L., Friedman H.S. Personality and Pathways of Influence on Physical Health // Social and Personality Psychology Compass. 2011. № 5. pp. 76–87. Doi: 10.1111%2Fj.1751-9004.2010.00331.x

[6] Wu W., Chen L., He L. Using personality to adjust diversity in recommender systems // HT 2013 - Proceedings of the 24th ACM Conference on Hypertext and Social Media. 2013. pp. 225–229. Doi: 10.1145/2481492.2481521.

[7] Kosinski M., Stillwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior // Proceedings of the National Academy of Sciences. 2013. № 15. pp. 5802–5806 Doi: 10.1073/pnas.1218772110

[8] Stankevich M., Smirnov I., Kiselnikova N., Ushakova A. Depression Detection from Social Media Profiles // Data Analytics and Management in Data Intensive Domains. 2022. pp. 181–194. Doi: 10.1007/978-3-030-51913-1_12

[9] Titov S., Novikov P., Mararitsa L. Full-scale Personality Prediction on VKontakte Social Network and its Applications // 2019 25th Conference of Open Innovations Association (FRUCT). 2019. pp. 317–323. Doi: 10.23919/FRUCT48121.2019.8981513.

[10] Ignatiev N., Smirnov I. and Stankevich M. Predicting Depression with Text, Image, and Profile Data from Social Media // International Conference on Pattern Recognition Applications and Methods. 2022. Doi: 10.5220/0010986100003122

[11] Grandini M., Bagli E., Visani G. Metrics for Multi-Class Classification: an Overview. 2020. Doi: 10.48550/ARXIV.2008.05756