

# Прогноз поведения человека на основе анализа данных социальных сетей

А.Н. Жданова<sup>1</sup>, А.В. Куприянов<sup>1</sup>, А.А. Хорошилова<sup>1</sup>

<sup>1</sup>Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе, 34а, Самара, Россия, 443086

## Аннотация

Статья посвящена решению проблемы сбора и анализа открытых данных пользователей социальных сетей с помощью методов машинного обучения. Такой подход позволяет прогнозировать поведение человека на основе анализа его сетевой активности. Для обработки неструктурированных данных были использованы нейронные сети TensorFlow и Keras, а также модели случайного леса и Gradient Booster. Результаты прогнозов были протестированы с использованием восьми различных моделей машинного обучения и протестированы на реальных данных, полученных в результате психологических и социологических исследований. Данный подход не зависит от входных данных и может применяться к любому набору данных.

## Ключевые слова

Прогнозирование, машинное обучение, социальная сеть, анализ публичных данных

## 1. Введение

В условиях роста популярности социальных сетей многие специалисты заинтересованы в возможности анализа открытых данных. Это открывает новые перспективы для исследований и решения бизнес-задач [1-3].

Целью исследования было создание прогностической модели интеллектуальных, эмоционально-волевых и коммуникативных составляющих личности, позволяющей предсказать реальное поведение субъекта на основе анализа открытой информации о его поведении в социальной сети.

## 2. Анализ публичных данных

Для исследования использовались результаты психологической диагностики: 1200 испытуемых за период 2007-2019 гг. – более 50 показателей, описывающих личностные, патопсихологические, характерологические и иные особенности испытуемых. В качестве источника публичных данных испытуемых была выбрана социальная сеть «ВКонтакте». Открытые данные профиля исследуемых были извлечены через API. На первом этапе были выбраны только количественные характеристики: пол, открытость профиля, количество друзей, подписчиков, фото и видео.

Для определения наиболее оптимального подхода предсказания характеристик были построены и исследованы следующие модели: линейная и логистическая регрессия, дерево, случайный лес, нейросети TensorFlow и Keras (с функцией активации Sigmoid и ReLu). Прогноз выдавался по каждому фактору в соответствии с классификацией: «высокий уровень», «норма», «низкий уровень».

Для повышения точности результатов работы моделей была использована 10-кратная перекрестная проверка, что позволило увеличить точность прогноза на 3-5%. Для проверки надежности предсказания характеристик использовался коэффициент Коэна (каппа Коэна).

В таблице 1 представлен результат работы каждой модели на выборке из 1200 человек 1990 – 1998 года рождения, где 206 из них – женщины, 905 – мужчины.

**Таблица 2**

Точность математических моделей

	Lin R	Log R	Tree	Forest	H2O Forest	H2O Gradient Booster	NN Keras	NN TensorFlow
A	64,71	73,53	76,47	47,06	76,47	35,29	47,06	50
B	47,06	67,65	48,42	50	73,53	8,23	58,82	50
C	44,17	62,94	58,82	44,12	67,65	23,53	47,06	52,94
E	57,06	62,94	73,53	44,12	83,53	20,59	52,94	55,88
F	62,27	60	58,2	31,82	80,46	36,36	38,64	47,73
G	69,09	53,18	64,8	27,27	78,18	22,73	47,73	50
H	84,1	20,59	72,42	23,53	68,82	29,41	47,06	44,12
I	79,47	65,88	62,2	55,88	11,76	26,47	47,06	44,12
L	64,71	73,53	76,47	47,06	76,47	35,29	47,06	50
M	47,06	67,65	48,42	50	73,53	8,23	58,82	50
N	44,17	62,94	58,82	44,12	67,65	23,53	47,06	52,94
O	57,06	62,94	73,53	44,12	83,53	20,59	52,94	55,88
Q1	62,27	60	58,2	31,82	80,46	36,36	38,64	47,73
Q2	69,09	53,18	64,8	27,27	78,18	22,73	47,73	50
Q3	84,1	20,59	72,42	23,53	68,82	29,41	47,06	44,12
Q4	79,47	65,88	62,2	55,88	11,76	26,47	47,06	44,12

Наилучшим образом проявили себя модели: H2O случайный лес, H2O Gradient Booster и грамотно подобранная модель нейросетей. Точность моделей – до 75-85%.

### 3. Заключение

Из полученных данных можно сделать вывод, что данный подход является перспективным. Некоторые психологические характеристики хорошо моделируются и подтверждаются несколькими рассмотренными методами. Таким образом, мы можем утверждать, что с помощью данного подхода возможно не только анализировать сетевое поведение субъекта, но и в определенной мере предсказывать его поведение в реальности.

### 4. Литература

- [1] Kosinski, M. Manifestations of user personality in website choice and behaviour on online social networks / M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, Th. Graepel. – Cambridge, USA, 2013. – 5 p.
- [2] Крылова (Смирнова), О.С. Описание информационного образа пользователя социальной сети с учетом его психологической характеристики / О.С. Крылова (Смирнова), Д.А. Власов // International Journal of Open Information Technologies. – 2018. – Vol. 6(4). – P. 24-37.
- [3] Maks, K. Predictive modeling in practice / K. Maks, D. Kell. – Springer, 2019. – 640 p.