Тенденции взаимосвязи личностных особенностей и результатов теста Голланда среди пользователей социальной сети ВКонтакте

Е. А. Глушков

Санкт-Петербургский государственный университет

В. Ф. Столярова

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

vfs@dscs.pro

Аннотация. **В силу особенностей существующих инструментов определения профориентационных предпочтений, в сфере профориентации возникает задача определения кода Голланда по альтернативным данным, как результаты психометрических тестов для валидации полученных результатов, а также использования в тех ситуациях, когда проведение теста затруднительно. Целью работы является формирование каскада моделей машинного обучения, позволяющего по результатам отдельных тестов и их комбинаций предсказывать принадлежность отдельного кода Голланда верхней или нижней триаде. В работе рассматриваются распространенные тесты: пятифакторный опросник личности, 16-факторный опросник Кеттелла, личностный опросник Айзенка, опросник Леонгарда–Шмишека, ценностный опросник Шварца. Были использованы данные 1278 респондентов, собранные при помощи приложения, размещенного на платформе VK Mini Apps. Для обработки данных применен метод главных компонент (уменьшение размерности) и многозадачной регрессии (модели градиентных бустингов, случайный лес, линейная регрессия, нейронные сети). Применены различные подходы решения задачи мультизадачного предсказания кода Голланда. Показано, что с точки зрения метрик RMSE и C-индекса наилучшие результаты показали модели Lasso-регрессии (RMSE = 1.914, C-index = 10.426), CatBoost (RMSE=1.904), случайный лес (C-индекс = 10.529) и ансамблевые модели. Разработанные модели составляют основу для инструмента определения социально-профессиональной направленности личности в случае отсутствия возможности прохождения самого теста, а также установления валидности результатов опроса согласно модели RIASEC без привязки к конкретному варианту теста.**

Ключевые слова: модель Голланда; RIASEC; социальные медиа; психологические тесты; многозадачная регрессия; ансамбль

I. Введение

Многие аспекты успешности человека обусловлены корректным определением карьерного пути, соответствующего его предпочтениям [1]. Карьерному самоопределению уделяются многие ресурсы, в том числе, и со стороны государства[1]. Золотым стандартом в сфере профориентации является глубинное интервью с экспертом, который поможет выявить сильные и слабые стороны личности. Однако этот подходявляется ресурсозатратным, и потому в качестве альтернативы развиваются дистанционные способы карьерного консультирования [2, 3]. Таким образом, актуальной является задача разработки моделей и алгоритмов выявления профессиональных предпочтений по доступным в дистанционном формате данным, таким как цифровые следы, профориентационные тесты [3–5].

Одним из инструментов для определения профессиональных интересов является модель RIASEC [6], которая предполагает наличие 6 типов социально-профессиональной направленности личности: реалистический, исследовательский, артистический, социальный, предприимчивый и конвенциональный. Существует множество вариаций тестов, которые отражают показатели этой шкалы [7], результаты которых коррелированы, однако не определяют друг друга однозначно. Кроме того, такие тесты часто не учитываются быстро изменяющуюся конъюнктуру рынка профессий, а также культурные и социо-экономические различия респондентов [7, 8]. Возникает актуальная задача определения кода Голланда по альтернативным данным, как результаты психометрических тестов для валидации полученных результатов, а также использования в тех ситуациях, когда проведение теста затруднительно. настоящее время проводятся исследования взаимосвязи между кодом Голланда и результатами «Большой Пятерки» [9–12] и цифровыми следами пользователей [4].

Целью работы является формирование каскада моделей машинного обучения, позволяющего по результатам отдельных психометрических тестов и их комбинаций предсказывать код Голланда. Разработанные модели позволяют определить конкретные значения кодов и предсказать, является ли тот или иной код ведущим (т. е. имеющим наибольшие значения) при определении социально-профессиональной направленности личности. Теоретическая значимость исследования состоит в выявлении взаимосвязи кода Голланда и результатов

Работа выполнена при финансовой поддержке в рамках проекта по государственному заданию СПБ ФИЦ РАН № FFZF-2025-0006

[1] Федеральный закон от 30 декабря 2020 г. № 489-ФЗ "О молодежной политике в Российской Федерации"

психометрических тестов, определении тех признаков, которые играют ключевую роль при определении социально-профессиональной направленности личности. Практическая значимость заключается в выборе модели или комбинации моделей машинного обучения для предсказания кода Голланда, что может служить основой для создания программного продукта оценки профессиональной направленности по психологическому профилю личности.

II. Материалы и методы

A. *Характеристика используемых тестов*

Социально-профессиональный тип личности согласно теории Дж. Голланда определяется как набор из шести факторов – реалистический (Realistic, R), исследовательский (Investigative, I), артистический (Artistic, A), социальный (Social, S), предприимчивый (Enterprising, E) или традиционный (Conventional, C), – которые выстраиваются в порядке убывания их выраженности. В исследовании использовалась адаптация методики Г. В. Резапкиной [13], которая предлагает респонденту 42 пары профессий, из которых респондент должен выбрать предпочтительную. Конкретное значение фактора определяется количеством выбранных профессий, которые относятся к определенному профессиональному типу. Таким образом, результатом прохождения теста является набор из 6 числовых значений, соответствующих кодам (факторам) RIASEC. Каждый фактор может принимать значения от 0 до 14, суммарный балл оценки всех факторов равен 42 (по количеству вопросов в адаптации опросного инструмента). В исследовании рассматривается также тройка наиболее выраженных кодов – верхняя триада.

Для сравнения профилей личности существуют различные меры конгруэнтности (согласованности), и для трехбуквенных кодов (верхних триад) используется C-индекс, рассчитываемый как

$$C = 3(X_1, Y_1) + 2(X_2, Y_2) + (X_3, Y_3),$$

где X_1, X_2, X_3 и Y_1, Y_2, Y_3 – триады кода Голланда, (X_i, Y_i) представляют собой различия между кодами Голланда, принимают значения от 0 до 3 в зависимости от того, совпадают ли коды (3 балла), соседствуют ли в замкнутой цепочке R-I-A-S-E-C напрямую (2) или через один код (1), иначе 0; чем больше значение C-индекса, тем более схожи профили личности.

В исследовании рассматривались также результаты тестов

- Опросник Леонгарда–Шмишека (10 факторов, идентификатор LN);
- Личностный опросник Айзенка (4 факторов, идентификатор EY);
- 16-факторный опросник Кеттелла (16 факторов, идентификатор CT);
- Пятифакторный опросник личности («Большая Пятерка») (5 факторов, идентификатор BF);
- Ценностный опросник Шварца (20 факторов, идентификатор SC).

Каждый из представленных факторов является числовой метрической переменной, принимающей только целочисленные значения.

B. *Используемые методы машинного обучения в задаче определения кода Голланда*

Предсказание значений кодов Голланда представляет собой предсказание вектора из шести чисел, соответствующих кодам RIASEC. В исследовании поставленная задача рассматривалась как задача регрессии со множественными выходами (многозадачной регрессии, англ. multioutput regression). Важное ограничение задачи заключается в том, что каждым из значений шести целевых переменных является целое число, при этом их сумма равна строго 42. Рассматривалось предсказание целевых переменных по цепочке (англ. chain regressor), когда каждое предсказанное значение целевой переменной становится предиктором для предсказания следующей переменной.

Предсказание по цепочке может учитывать связи между целевыми переменными, однако для этого требуется больше данных, неправильный порядок может ухудшить качество. В качестве базовых моделей для независимого предсказания и предсказания по цепочке были использованы следующие регрессионные модели:

- линейная регрессия (в т. ч. с L1- и L2-регуляризацией);
- градиентные бустинги (XGBoost, LightGBM, CatBoost);
- случайный лес;
- метод k-ближайших соседей;
- метод опорных векторов (регрессор);
- многослойный перцептрон (3 скрытых слоя: 128, 64 и 32 нейрона, функция активации ReLU, один выходной слой с 6 нейронами, оптимизатор Adam);
- foundation-модель для табличных данных (TabPFN).

Для улучшения качества прогноза также было использовано взвешенное ансамблирование моделей, где итоговый ответ вычислялся как линейная комбинация предсказаний различных моделей с оптимизированными коэффициентами.

Для оценки качества моделей были использованы следующие метрики: усредненное по всем объектам значение среднеквадратичной ошибки (RMSE), усредненное значение C-индекса [14].

В работе [14] рассматривалась схожая задача предсказания кода Голланда на основе социально-демографических признаков. Авторы предлагают различные подходы для решения этой задачи: с тех пор как код Голланда может быть представлен и как последовательно идущие 3 или 6 букв, и как значения, соответствующие кодам, то и задача может быть поставлена следующим образом: многоцелевая регрессия (multioutput regression), классификация с несколькими метками (multilabel classification), многоцелевая классификация (multioutput classification).

В статье отмечается: в случае последовательного предсказания для многоцелевой регрессии порядок предсказания выходов важен. В качестве метрики авторы используют меру конгруэнтности – C-индекс, для уменьшения размерности – метод главных компонент (PCA). Лучшие результаты показал градиентный бустинг: как при решении задачи регрессии, так и при задаче классификации $C = 11.08$ (для сравнения: при случайном ответе $C = 9$).

C. Используемые данные

Данные для исследования были собраны с помощью веб-приложения на базе платформы VK Mini Apps «Психологические тесты»[2], которое размещено в свободном доступе. При этом после ознакомления с условиями добровольного информированного согласия пользователи могут разрешить использовать обезличенные анонимизированные данные в научных исследованиях (в соответствии с № 152-ФЗ «О персональных данных»). Для исследования была взята выборка из 1278 респондентов, прошедших упомянутые психометрические тесты. Для 339 опрашиваемых, данные заполнены по всем 6 тестам, по 939 имеются данные лишь по 4 или 5 различным тестам.

III. Моделирование и результаты исследования

Статистическое моделирование согласно описанной методологии проводилось с помощью языков программирования R (версия 4.4.2) и Python (для перцептрона и TabPFN; версия 3.12.7).

Предобработка данных включает приведение к широкому формату; были заполнены пропуски для данных по тестам при наличии заполненных данных по другим факторам теста; выполнены проверки ограничений на допустимые значения факторов. Также определялась коррелированность наблюдений, из каждого набора коррелированных в сильной степени факторов ($r > 0.7$) оставался лишь один признак.

Моделирование включало построение моделей на полных данных, включающих в себя все факторы всех психометрических тестов, и моделей на данных по отдельным психометрическим тестам. Кроме того, на примере LightGBM был проведен анализ важности признаков как общий прирост функции разделения: были найдены средние значения важности каждого из признаков по всем шести целевым переменным.

Оценка качества проводилась на контрольном (тестовом) наборе данных.

В табл. 1 приведены результаты предсказания кода Голланда на полных данных (значения метрик RMSE и C-индекс) без и со снижением размерности методом PCA. В табл. 2 и 3 представлены усредненные значения метрик RMSE и C-индекса на данных по отдельным психометрическим тестам. Оценка важности признаков-предикторов (усредненная по всем шести целевым признакам) для модели LightGBM на полных данных приведена в табл. 4.

В таблицах используется также цветовое обозначение, зеленым цветом обозначены лучшие решения (минимум RMSE и максимум C-индекса), красным – наихудшие решения.

ТАБЛИЦА I. Результаты предсказания кода Голланда на полных данных на основе метрики RMSE

Модель	RMSE		C-индекс	
	Полные данные	PCA	Полные данные	PCA
Базовая константная	2.308	2.308	8.94	8.94
Линейная регрессия	2.122	1.993	9.706	10.265
Lasso-регрессия (L1)	**1.916**	**1.914**	**10.235**	**10.426**
Гребневая регрессия (L2)	1.933	**1.918**	9.5	9.956
XGBoost	2.265	2.102	9.103	10.118
LightGBM	1.964	1.963	**10.324**	**10.368**
CatBoost	**1.904**	1.943	9.338	10.103
Случайный лес (Random Forest)	1.932	1.971	9.324	**10.529**
k-ближайших соседей (kNN)	2.036	2.001	9.015	9.779
Метод опорных векторов (SVR)	2.001	2.001	9.603	**10.368**
Многослойный перцептрон	2.11	-	9.401	-
TabPFN	2.217	-	8.98	-
Ансамбль 1[1]	1.935	1.95	9.632	9.926
Ансамбль 2[2]	**1.913**	1.952	**10.132**	10.001
Ансамбль 3[3]	**1.911**	**1.919**	**10.118**	9.897

Примечания:
Ансамбль 1 – все указанные в таблице модели, кроме TabPFN, перцептрона и ансамблей;
Ансамбль 2 – Случайный лес, LightGBM, kNN (с весами 0.636, 0.273, 0.091);
Ансамбль 3 – CatBoost, Lasso-регрессия (L1), LightGBM, Случайный лес, Гребневая регрессия (L2) (с весами 0.6, 0.25, 0.05, 0.05, 0.05).

ТАБЛИЦА II. Результаты предсказания кода Голланда (RMSE) на данных по отдельным психометрическим тестам

Модель	BF	CT	EY	SC	LN
Базовая константная	2.229	2.230	2.230	2.230	2.232
Линейная регрессия	2.157	2.054	2.173	2.183	2.151
XGBoost	2.403	2.184	2.298	2.373	2.329
LightGBM	2.208	2.106	2.202	2.210	2.159
CatBoost	2.188	2.069	2.182	2.180	2.143
Случайный лес (Random Forest)	2.233	2.083	2.223	2.194	2.170
k-ближайших соседей (kNN)	2.265	2.150	2.271	2.269	2.224
Метод опорных векторов (SVR)	2.261	2.135	2.246	2.236	2.234
Lasso-регрессия (L1)	2.186	2.279	2.127	2.157	2.141
Гребневая регрессия (L2)	2.193	2.276	2.126	2.167	2.142

ТАБЛИЦА III. Результаты предсказания кода Голланда (C-индекс) на данных по отдельным психометрическим тестам

Модель	BF	CT	EY	SC	LN
Базовая константная	9.799	9.799	9.466	9.799	9.466
Линейная регрессия	9.419	9.879	9.540	9.861	9.658
XGBoost	9.236	9.558	9.696	9.702	9.460
LightGBM	9.395	9.976	9.578	9.451	9.894
CatBoost	9.687	10.109	9.369	9.885	9.864
Случайный лес (Random Forest)	9.333	10.056	9.649	9.761	9.342
k-ближайших соседей (kNN)	9.159	9.599	9.528	9.416	9.204
Метод опорных векторов (SVR)	9.552	9.941	9.516	9.611	9.572

[2] Мини-приложение "Психологические тесты" (платформа "VK Mini Apps") [Электронный ресурс]. URL: https://vk.com/app7794698 (дата обращения: 27.03.2025)

Модель	BF	CT	EY	SC	LN
Lasso-регрессия (L1)	9.676	9.956	10.029	8.309	9.103
Гребневая регрессия (L2)	9.794	9.941	9.706	8.338	9.029

ТАБЛИЦА IV. УСРЕДНЕННАЯ ОЦЕНКА ВАЖНОСТИ ПРЕДИКТОРОВ ДЛЯ МОДЕЛИ LIGHTGBM

Код признака	Наименование признака	Важность признака (%)	Накопленное значение важности (%)
CT_1	Открытость - Замкнутость	14.9	14.9
CT_7	Чувственность - Твердость	6.9	21.8
SC_1	Безопасность - индивидуальный приоритет	4.5	26.3
CT_10	Утонченность - Простота	4.5	30.8
EY_1	Экстраверсия	3.9	34.6
SC_19	Традиция - нормативный идеал	2.6	37.2
SC_5	Достижение - индивидуальный приоритет	2.4	39.6
SC_3	Гедонизм - индивидуальный приоритет	2.3	41.9
BF_5	Экспрессивность - практичность	2.3	44.2
LN_1	Гипертимность	2.3	46.5
BF_1	Экстраверсия - интроверсия	2.2	48.7
CT_3	Независимость - Податливость	2.2	50.8

IV. ОБСУЖДЕНИЕ

На полном наборе всех тестов наилучшие результаты (табл. 1) показали модели CatBoost (RMSE = 1.904), Lasso-регрессии (L1) с использованием PCA (RMSE = 1.914, C-индекс = 10.426), ансамблевые модели, в частности комбинация моделей CatBoost, Lasso-регрессии, LightGBM, случайного леса и гребневой регрессии (RMSE = 1.911, C-индекс = 10.118), случайный лес с PCA (C-индекс = 10.529). Наиболее стабильные результаты сразу по обеим метрикам показала модель Lasso-регрессии с PCA. Отметим, что результаты всех моделей оказались лучше, чем значения метрик базовой константной модели, предсказания которой были всегда равны средним по целевым переменным на обучающем наборе данных.

По результатам предсказания целевых переменных только лишь на данных отдельных тестов (табл. 2 и 3) наибольшие значения метрик достигаются на данных 16-факторного опросника Кеттелла и опросника Шварца (в меньшей степени), что может быть связано с тем, что именно они имеют наибольшее число факторов (16 и 20 соответственно). Наилучшие значения метрик достигаются при использовании моделей CatBoost и случайного леса, хуже всего – XGBoost и метод k-ближайших соседей.

Согласно табл. 4 для модели LightGBM наиболее важными факторами являются три фактора опросника Кеттелла – открытость, чувственность, утонченность, – а также фактор «Безопасность» (опросник Шварца). Данные четыре показателя обеспечивают 30 % важности от всех 55 исходных факторов.

К основным ограничениям исследования относятся особенности сбора данных: возможны смещения из-за специфики портала, а также способа формирования выборки. Для устранения ограничений может быть увеличен размер выборки, включены вопросы о социально-демографических признаках опрашиваемых.

V. ЗАКЛЮЧЕНИЕ

Таким образом, был сформирован каскад моделей машинного обучения, позволяющий по результатам отдельных психометрических тестов и их комбинаций предсказывать код Голланда, обеспечивающий наилучшие значения выбранных метрик (RMSE и C-индекса). Наилучшие результаты показали модели Lasso-регрессии (RMSE = 1.914, C-index = 10.426), CatBoost (RMSE = 1.904), случайный лес (C-индекс = 10.529) и ансамблевые модели на полных данных и модели CatBoost и случайный лес на данных по тестам отдельно. В качестве наиболее важных предикторов можно выделить факторы «Открытость» и «Чувственность» (оба – 16-факторный опросник Кеттелла). Выбор моделей (каскада моделей) для предсказания кода Голланда может служить основой для автоматизации оценки профессиональных предпочтений индивида, а именно для создания программного продукта оценки профессиональной направленности по психологическому профилю личности.

СПИСОК ЛИТЕРАТУРЫ

[1] Presti A. L. et al. Career competencies and career success: On the roles of employability activities and academic satisfaction during the school-to-work transition //Journal of Career Development. 2022. Т. 49. № 1. С. 107-125.

[2] Pordelan N., Hosseinian S. Design and development of the online career counselling: a tool for better career decision-making //Behaviour & Information Technology. 2022. Т. 41. №. 1. С. 118-138.

[3] Westman S. et al. Artificial Intelligence for Career Guidance–Current Requirements and Prospects for the Future //IAFOR Journal of Education. 2021. Т. 9. №. 4. С. 43-62.

[4] Chekalev A., Khlobystova A., Abramov M. Community Theme Analyser: Predicting Career Guidance in Online Social Networks //International Conference on Intelligent Information Technologies for Industry. Cham : Springer Nature Switzerland, 2024. С. 153-162.

[5] Oliseenko V., Ivaschenko A., Korepanova, A. Tulupyeva T. Automating the Temperament Assessment of Online Social Network Users // Doklady Mathematics. 2024. Vol. 108. С. 368-373. DOI:10.1134/S1064562423701041

[6] Holland, J. L (1985). Making vocational choices: A theory of vocational personalities and work envronments. Prentice-Hall.

[7] Chu C. et al. What do interest inventories measure? The convergence and content validity of four RIASEC inventories //Journal of Career Assessment. 2022. Т. 30. №. 4. С. 776-801.

[8] Hoff K. A., Granillo-Velasquez K. E., Hanna A., Morris M., Nelson H., Oswald F. L. Interested and employed? A national study of gender differences in basic interests and employment. Journal of Vocational Behavior. 2024. DOI: 10.1016/j.jvb.2023.103942

[9] Mason R., Roodenburg J., Williams B. Personality and vocational interest typologies associated with better coping and resilience in paramedicine: A review of two models //Paramedicine. 2024. Т. 21. № 1. С. 36-44.

[10] Hurtado Rúa S. M., Stead G. B., Poklar A. E. Five-factor personality traits and RIASEC interest types: A multivariate meta-analysis //Journal of Career Assessment. 2019. Т. 27. №. 3. С. 527-543.

[11] Schuerger J. M. Career assessment and the sixteen personality factor questionnaire //Journal of Career Assessment. 1995. Т. 3. № 2. С. 157-175.

[12] Yamashita J. et al. Personality traits systematically explain the semantic arrangement of occupational preferences //Journal of Individual Differences. 2024. 45(4). С. 201–217.

[13] Резапкина Г.В. Психология и выбор профессии: программа предпрофильной подготовки. Учебно–методическое пособие для психологов и педагогов. М.: Генезис, 2005.

[14] Bogacheva E., Tatarenko F., Smetannikov I. Predicting vocational personality type from socio-demographic features using machine learning methods //Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System. 2020. С. 93-98.