



Санкт-Петербургский государственный университет
Кафедра информатики

Определение кода Голланда по результатам психометрических тестов личности на основе методов машинного обучения в условиях неполноты информации

ГЛУШКОВ Егор Александрович

Научный руководитель: доцент кафедры информатики, к. т. н., Абрамов М. В.

Консультант: ст. преподаватель кафедры информатики, Столярова В. Ф.

Рецензент: ст. научный сотрудник, СПб ФИЦ РАН, к. т. н., Захаров В. В.

Санкт-Петербург, 2025

- Важность корректного выбора профессионального пути
- Ресурсоёмкость традиционного глубинного интервью, дистанционные методы профориентации, психометрические тесты
- Модель RIASEC (код Голланда):
 - ▶ шесть типов социально-профессиональной направленности личности
 - ▶ различные вариации теста
 - ▶ сравнение профессиональных профилей с помощью С-индекса
- Взаимосвязь кода Голланда с социально-демографическими признаками, цифровыми следами, психометрическими тестами с помощью статистических методов, структурного моделирования, машинного обучения
- Нет инструментов, позволяющих по комбинации популярных тестов предсказывать код Голланда, особенно в условиях неполноты информации

Постановка задачи

Целью работы является автоматизация процесса профориентации посредством разработки инструмента для предсказания кода Голланда по неполным результатам психометрических тестов с использованием методов машинного обучения

Задачи:

- Реализовать различные подходы к определению кода Голланда: многоцелевая регрессия, классификация, ранжирование
- Разработать модуль формирования взвешенного ансамбля моделей для объединения прогнозов базовых моделей
- Провести сравнительный анализ подходов и методов определения кода Голланда на основе С-индекса
- Разработать математическое обеспечение для модуля восстановления пропусков психометрических тестов
- Создать прототип инструмента для определения профориентационных предпочтений

- *Новизна результатов исследования:* создание нового программного комплекса, обеспечивающего автоматизацию процесса профориентации на основе предсказания кода Голланда
- *Теоретическая значимость:* использование уникальной комбинации различных психометрических тестов при разработке новых моделей машинного обучения для определения взаимосвязи тестов и кода Голланда
- *Практическая значимость:* разработка прототипа программного модуля автоматизации оценки профессиональной направленности по психологическому профилю личности

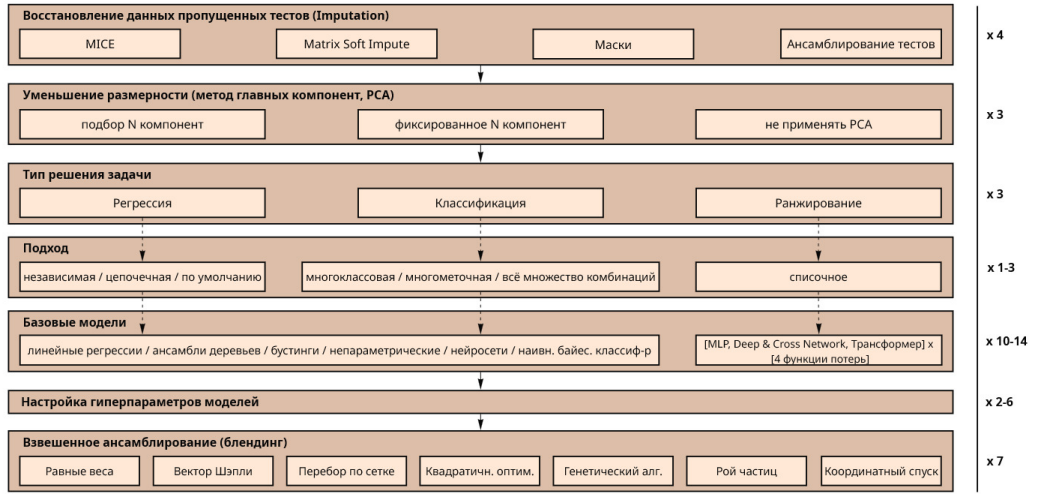
Обзор. Психометрические тесты личности

- Тест Голланда RIASEC (6)
6 типов (профилей) личностей, с которыми соотнесены наборы профессий
- Опросник Леонгарда-Шмишека (10)
- Личностный опросник Айзенка (4)
- 16-факторный опросник Кеттелла (16)
- Пятифакторный опросник личности («Большая пятерка»; 5)
- Ценностный опросник Шварца (20)

Таблица 1. Пример данных психометрических тестов

id	Большая пятёрка			...	Леонгард		Голланд					
	BF1	BF2	BF3		LN9	LN10	R	I	A	S	E	C
1	39	66	33	...	3	12	8	8	6	8	1	11
2	45	46	73	...	12	6	3	7	7	8	10	7
3	34	41	56	...	18	12	10	10	3	11	7	1
4	49	47	50	...	15	24	6	4	8	6	7	11

Общая схема вариантов вычислительного эксперимента



Итого 10-63 тыс. вариантов

Рисунок 1. Общая схема вариантов вычислительного эксперимента

Подходы и метрики качества

- Многоцелевая регрессия — метрика $avgRMSE$
- Классификация — $Top-k accuracy$
- Ранжирование — $NDCG@3$
- Сравнение подходов на основе C-индекса, желаемое значение: $C_{index} \geq 11$

$$C_{index} = 3(X_1, Y_1) + 2(X_2, Y_2) + 1(X_3, Y_3),$$

где $\{X_i\}$ и $\{Y_i\}$ — первые три позиции кодов Голланда, их позиции в замкнутой цепочке (шестиугольнике) $R-I-A-S-E-C$:

$$(X_i, Y_i) = \begin{cases} 3, & \text{если } X_i = Y_i, \\ 2, & \text{если } X_i \text{ и } Y_i - \text{соседние позиции,} \\ 1, & \text{если } X_i \text{ и } Y_i - \text{позиции через один код,} \\ 0, & \text{если } X_i \text{ и } Y_i - \text{противоположны.} \end{cases}$$

Особенности реализации

- R (версия 4.4.2):
 - ▶ векторизованная обработка и манипуляции с данными: *data.table*, *tidyverse*, *R6*
 - ▶ статистические и ML-модели: *stats*, *mice*, *softImpute*, *glmnet*, *MASS*, *xgboost*, *lightgbm*, *catboost*, *randomForest*, *FNN*, *caret*, *e1071*, *ranger*, *quadprog*, *GA*, *PSO*
 - ▶ интерактивные веб-приложения: *Shiny* (*Posit*); визуализация: *plotly*
- Python (версия 3.12.3):
 - ▶ *numpy*, *pandas*, *sklearn*, *PyTorch*, *TabPFN*
- Основные этапы реализации¹:
 - 1 Разведочный анализ данных
 - 2 Вычислительный эксперимент: выбор наилучшего подхода к определению кода Голланда, обучение и сохранение модели
 - 3 Прототип инструмента автоматизации профориентации (веб-приложение)

¹ GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности.

URL: https://github.com/Exp98/Diploma_Holland (дата обращения: 07.06.2025)

Архитектура вычислительного модуля

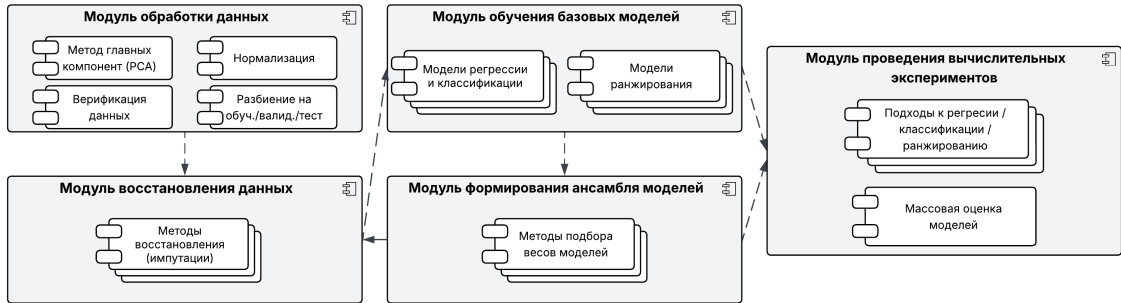


Рисунок 2. Архитектура вычислительного модуля

Описание набора данных

- VK Mini Apps «Психологические тесты»²
- Анонимизированные данные³ 1278 пользователей: 339 — полные, 939 — не заполнены данные по 1–2 тестам
- Обработка данных: json → широкий табличный формат, валидация, заполнение пропусков, нормализация, понижение размерности (метод главных компонент, *PCA*)
- Ограничения: особенности сбора данных (смещения из-за специфики портала, способа формирования выборки)

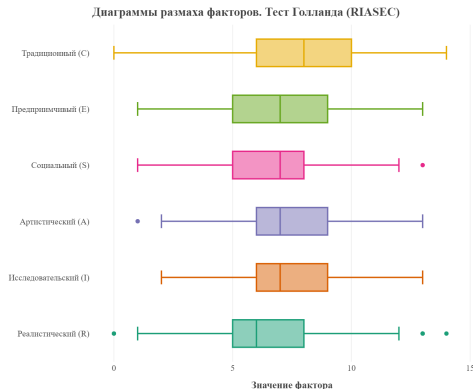


Рисунок 3. Диаграмма размаха факторов Голланда

² Мини-приложение «Психологические тесты» («VK Mini Apps»). URL: <https://vk.com/app7794698>

³ Политика конфиденциальности. URL: <https://vk.com/@ticslabs-politika-konfidencialnosti>

- Регрессия и классификация:

- ▶ модели на основе линейной регрессии (*Lasso L1, Ridge L2, пошаговая*)
- ▶ ансамбли деревьев (*случайный лес, ExtraTrees*)
- ▶ модели градиентного бустинга (*CatBoost, LightGBM, XGBoost*)
- ▶ непараметрические модели (*kNN, метод опорных векторов*)
- ▶ нейросетевые модели (*MLP, TabPFN*)
- ▶ регуляризованная логистическая регрессия (*Lasso L1, Ridge L2*)
- ▶ наивный байесовский классификатор
- ▶ базовая константная модель (для сравнительного анализа)

- Списочное ранжирование:

- ▶ Скоринговая функция: MLP, Deep & Cross Network, трансформер
- ▶ Функции потерь: ApproxNDCG, LambdaRank, ListNet@1, ListNet@3

Сравнение регрессионных моделей

Таблица 2. Сравнение базовых регрессионных моделей по C-индексу

Модель	Multioutput		Chained	
	без PCA	PCA	без PCA	PCA
Регрессия Lasso (L1)	11.175	10.887	11.175	11.150
ExtraTrees	10.700	11.100	10.625	10.825
Регрессия Ridge (L2)	10.988	10.537	11.062	10.412
Метод опорных векторов	10.713	10.950	10.713	10.950
Пошаговая регрессия	10.605	10.905	10.600	10.905
CatBoost	10.688	10.812	10.688	10.812
Случайный лес	10.625	10.475	10.812	10.588
LightGBM	10.750	10.425	10.750	10.425
k-ближайших соседей (kNN)	10.525	10.400	10.525	10.400
XGBoost	9.164	9.729	9.162	9.725
Базовая константная	9.000	9.000	9.000	9.000
TabPFN	10.562			
MLP (BN, DropOut, регуляризация)	10.462			
Многослойный перцептрон (MLP)	10.275			

Сравнение подходов к классификации

Таблица 3. Сравнение подходов к классификации (метрика Top-K accuracy)

Модель	Multiclass			Multilabel			Label Powerset		
	Top1	Top2	Top3	Top1	Top2	Top3	Top1	Top2	Top3
kNN	0.99	0.71	0.13	1.00	0.76	0.11	0.98	0.65	0.18
Логистич. L1-регр.	1.00	0.70	0.16	1.00	0.70	0.16	0.99	0.64	0.10
XGBoost	1.00	0.70	0.11	0.98	0.68	0.10	0.96	0.63	0.11
Логистич. L2-регр.	1.00	0.70	0.15	0.99	0.70	0.21	0.99	0.68	0.09
Наивный Байес	0.98	0.70	0.15	0.99	0.70	0.15	0.99	0.69	0.16
ExtraTrees	1.00	0.73	0.15	1.00	0.78	0.15	0.98	0.69	0.20
Метод опорн. вект.	1.00	0.74	0.15	1.00	0.72	0.14	0.98	0.68	0.21
Случайный лес	1.00	0.74	0.16	1.00	0.74	0.15	0.99	0.64	0.23
CatBoost	0.99	0.79	0.11	0.99	0.79	0.11	0.99	0.70	0.16
LightGBM	0.98	0.66	0.09	0.98	0.70	0.10	0.95	0.63	0.09

Сравнение классификационных моделей

Таблица 4. Сравнение базовых классификационных моделей

Классификатор	Подход	С-индекс	Top1	Top2	Top3
k-ближайших соседей (kNN)	Multilabel	10.838	1.000	0.763	0.113
Логистическая L1-регрессия	Multiclass	10.663	1.000	0.700	0.163
XGBoost	Multiclass	10.638	1.000	0.700	0.113
Логистическая L2-регрессия	Multiclass	10.500	1.000	0.700	0.150
Наивный байесовский класс-р	Multilabel	10.350	0.988	0.700	0.150
ExtraTrees	Multilabel	10.013	1.000	0.775	0.146
Метод опорных векторов	Multilabel	9.875	1.000	0.721	0.138
Случайный лес	Multilabel	9.800	0.996	0.738	0.146
CatBoost	Multilabel	9.775	0.988	0.788	0.113
LightGBM	Multilabel	9.313	0.975	0.700	0.100
Базовый случайный	—	9.000	0.950	0.500	0.050

Сравнение моделей ранжирования

Таблица 5. Сравнение моделей ранжирования

Функция потерь	С-индекс			NDCG@3		
	Deep&Cross	Transformer	MLP	Deep&Cross	Transformer	MLP
ApproxNDCG	10.025	8.888	9.150	0.539	0.439	0.388
LambdaRank	9.963	9.675	9.650	0.527	0.489	0.543
ListNet@1	9.650	10.325	10.438	0.504	0.628	0.653
ListNet@3	9.450	9.950	10.788	0.458	0.622	0.638

Ансамблирование регрессионных моделей

Таблица 6. Сравнение методов подбора весов ансамбля регрессионных моделей

Метод подбора весов	Multioutput		Chained		
	все модели	топ-5	все модели	топ-5	
Равные веса всех моделей	11.063	11.088	11.050	11.013	
Вектор Шэпли (Shar)	11.050	11.138	11.138	11.050	
Частичный перебор по сетке	11.550	11.388	11.538	11.325	
Квадратичная оптимизация (QP)	10.588	10.463	10.738	10.813	
Генетический алгоритм (GA)	11.500	11.550	11.300	11.563	
Метод роя частиц (PSO)	11.600	11.663	11.613	11.613	
Координатный спуск	11.188	11.225	11.288	11.413	
Лин. регрессии с регуляризацией L1, L2, LightGBM, CatBoost, RF	Линейная регрессия		10.887		
	Линейная регрессия		10.688		
Подбор весов	Lasso L1	Пошаговая перп.	CatBoost	ExtraTrees	Σ
PSO, топ-5 multioutput	0.432	0.327	0.150	0.091	= 1.000

Ансамблирование классификационных моделей

Таблица 7. Сравнение методов подбора весов ансамбля классификаторов

Метод подбора весов	Multiclass	Multilabel	Label Powerset
Равные веса всех моделей	10.663	10.888	10.563
Вектор Шэпли (Shap)	10.563	11.038	10.525
Частичный перебор по сетке	11.213	11.488	11.525
Квадратичная оптимизация (QP)	10.488	10.638	10.650
Генетический алгоритм (GA)	11.263	11.313	11.213
Метод роя частиц (PSO)	11.263	11.625	11.525
Координатный спуск	11.200	11.275	10.425

Таблица 8. Весовые коэффициенты моделей для PSO

Подбор весов	kNN	Logit L1	XGBoost	SVM	LightGBM	...	Σ
PSO	0.291	0.191	0.183	0.164	0.151	0.020	= 1.000

Методы восстановления результатов незаполненных тестов

Таблица 9. Восстановление значений незаполненных психометрических тестов с помощью базовых регрессионных моделей

Модель-регрессор	MICE	Soft Impute	Маски	Ансамбли
Регрессия Lasso (L1)	9.191	10.248	9.998	9.866
Пошаговая регрессия	9.608	10.183	9.978	10.082
Случайный лес	9.518	10.136	9.819	9.712
LightGBM	9.372	10.086	9.686	9.594
Линейная регрессия (OLS)	9.407	10.021	9.876	10.012
Регрессия Ridge (L2)	9.442	9.770	9.868	9.933
ExtraTrees	9.101	9.823	9.870	9.808
Метод опорных векторов	9.221	9.814	9.864	9.760
CatBoost	9.131	9.814	9.835	9.461
k-ближайших соседей (kNN)	9.372	9.834	9.830	9.377
XGBoost	8.769	9.571	9.267	9.614
Базовая константная	9.000	9.000	9.000	9.000

Ансамблевые методы восстановления для Soft Impute

Таблица 10. Весовые коэффициенты моделей и С-индекс при разных методах подбора весов для ансамблей на восстановленных данных

Подбор весов	Веса моделей					С-индекс
	Lasso L1	Пошагов.	LightGBM	Случ. лес	kNN	
Метод роя частиц (PSO)	0.001	0.481	0.038	0.475	0.005	10.740
Частичный перебор по сетке	0.000	0.500	0.000	0.500	0.000	10.657
Генетический алгоритм (GA)	0.281	0.369	0.109	0.189	0.052	10.401
Координатный спуск	0.019	0.422	0.067	0.305	0.187	10.245
Квадратичная оптимиз. (QP)	0.050	0.000	0.390	0.007	0.553	10.065
Равные веса всех моделей	0.200	0.200	0.200	0.200	0.200	10.053
Вектор Шэпли (Shar)	0.247	0.185	0.206	0.179	0.183	10.047

Обозначения:

Пошагов. — пошаговая регрессия,

Случ. лес — случайный лес (Random Forest),

kNN — метод k -ближайших соседей

Итоговая последовательность вычислительных шагов

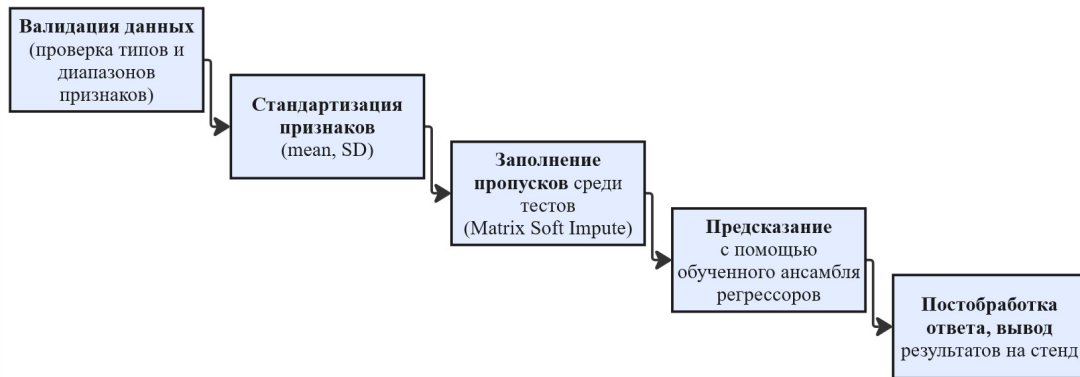


Рисунок 4. Итоговая последовательность шагов вычислительного конвейера⁴

⁴ Предсказание кода Голланда по результатам психометрических тестов — Shinyapps.io
URL: https://exp98.shinyapps.io/diploma_holland (дата обращения: 07.06.2025)

Интерфейс прототипа инструмента профориентации

Предсказание кода Голланда по результатам психометрических тестов

☐ Тест 16-факторный опросник Кеттелла (16 факторов)

☐ Тест Личностный опросник Айзенка (4 фактора)

☐ Тест Опросник Леонгарда-Шмишека (10 факторов)

☒ Тест Пятифакторный опросник личности (5 факторов)

1. Экстраверсия - интроверсия

43

Допустимо: от 15 до 75

2. Привязанность - обособленность

50

Допустимо: от 15 до 75

3. Самоконтроль - импульсивность

51

Допустимо: от 15 до 75

4. Эмоциональная устойчивость - неустойчивость

54

Допустимо: от 15 до 75

5. Экспрессивность - практичность

55

Допустимо: от 15 до 75

☐ Тест Ценностный опросник Шварца (20 факторов)

Подсчитать

Результаты прогноза

Прогноз сделан на основе результатов следующих тестов:

- Пятифакторный опросник личности

Коды Голланда:

- Наиболее вероятные: I (50.4%), C (18%), R (16.3%)
- Менее вероятные: S (10.5%), A (2.8%), E (2%)


Обозначения:

X (Y%), где X - код Голланда, соответствующий типу личности,

Y - степень уверенности, что данный код Голланда входит в верхнюю триаду

Ваши типы личности:

- I (Исследовательский).

Ловит анализировать данные, исследовать гипотезы и решать интеллектуальные задачи. Стремится к научным открытиям и пониманию сложных систем. Примеры: учёный , программист, биолог, химик.

- C (Конвенциональный).

Предпочитает чёткие инструкции, структуру и работу с цифрами/документами. Ценит аккуратность и системный подход. Примеры: бухгалтер , архивариус, налоговый инспектор, логист.

- R (Реалистичный).

Предпочитает практические задачи, работу руками и с техникой. Часто выбирает профессии, связанные с физическим трудом или природой. Примеры: инженер , механик, строитель, фермер.

Подсчет выполнен

×

Результаты экспериментов

- Лучшая базовая модель — L1-регрессия с независимыми выходами:
 $C_{index} = 11.175$
- Превосходство классических методов машинного обучения над нейросетевыми — лучшая среди нейронных сетей MLP (ListNet@3): $C_{index} = 10.788$
- Ансамблевые модели (подбор весов методом роя частиц):
 - ▶ ансамбль регрессоров с независимыми выходами: $C_{index} = 11.663$
 - ▶ ансамбль классификаторов со многими метками: $C_{index} = 11.625$
- Для восстановления данных предпочтителен метод мягкой импутации *Soft Impute*: $C_{index} = 10.740$

Результаты

- ❶ Реализованы⁵ подходы к определению кода Голланда: регрессия, классификация, списочное ранжирование
- ❷ Разработан модуль формирования взвешенного ансамбля моделей
- ❸ Проведен сравнительный анализ моделей предсказания кодов Голланда
- ❹ Разработаны и реализованы математические модели модуля восстановления пропусков результатов психометрических тестов
- ❺ Создан прототип инструмента для определения профориентационных предпочтений на основе R Shiny
 - Участие в XXVIII Международной конференции SCM'25
 - Акт об использовании результатов ВКР в НИР СПб ФИЦ РАН

⁵ GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности.
URL: https://github.com/Exp98/Diploma_Holland (дата обращения: 07.06.2025)

Акт об использовании результатов ВКР

МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
«САНКТ-ПЕТЕРБУРГСКИЙ ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
РОССИЙСКОЙ АКАДЕМИИ НАУК» (СПб ФИЦ РАН)

14-я линия В.О., д. 39, Санкт-Петербург, 199178

Телефон: (812) 508-33-11, факс: (812) 328-44-50, Email: info@spcras.ru, https://spcras.ru/
ОКПО:04683303, ОГРН:1027800514411, ИНН/КПП:7801003920/780101001

УТВЕРЖДАЮ

заместитель директора
СПб ФИЦ РАН
по безопасности

В.С. Поляков

«7» июня 2025

А К Т

Об использовании результатов выпускной квалификационной работы
Глушкова Егора Александровича
«Определение кода Голланда по результатам психометрических тестов
личности на основе методов машинного обучения в условиях неполноты
информации»

в научно-исследовательской работе СПб ФИЦ РАН

Комиссия в составе: председателя – в.н.с., к.т.н. Абрамова М.В., членов
комиссии: и.с. Олисеенко В.Д., м.н.с. Столяровой В.Ф. составила настоящий акт
о том, что результаты, полученные Глушковым Егором Александровичем в
процессе выполнения выпускной квалификационной работы «Определение кода
Голланда по результатам психометрических тестов личности на основе методов
машинного обучения в условиях неполноты информации», в том числе:

- прототип инструмента для определения профориентационных предпочтений личности
- математическое обеспечение определения кода Голланда по неполному набору результатов психометрических тестов личности (модель восстановления пропусков в данных, формирования взвешенного ансамбля моделей для объединения прогнозов базовых алгоритмов)

были внедрены в научно-исследовательской работе СПб ФИЦ РАН, FFZF-2024-0003 «Теоретические и технологические основы персонализированного обслуживания пользователей финансово-технических и социоориентированных систем с применением генеративных моделей обработки мультимодальных данных и цифровых следов». Полученные в ходе работы модели и методы анализа направлены на выявление и сопоставление психологических и поведенческих

характеристик обслуживающего лица и клиента для оптимизации взаимодействия и повышения эффективности обслуживания и будут использованы для развития прототипа программного продукта для формирования рекомендаций при обслуживании клиентов.

Председатель комиссии:

Руководитель НИР, в.н.с.,
к.т.н.,
Абрамов Максим Викторович

Члены комиссии:

Исполнитель НИР,
и.с. Олисеенко Валерий Дмитриевич

Исполнитель НИР,
м.н.с. Столярова Валерия Фуатовна



Ссылки на дополнительные материалы

- Исходный код¹
- Веб-приложение²
- Участие в XXVIII Международной конференции по мягким вычислениям и измерениям SCM'25³

¹ GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности. URL: https://github.com/Exp98/Diploma_Holland (дата обращения: 07.06.2025)

² Предсказание кода Голланда по результатам психометрических тестов - Shinyapps.io. URL: https://exp98.shinyapps.io/diploma_holland (дата обращения: 07.06.2025)

³ Тенденции взаимосвязи личностных особенностей и результатов теста Голланда среди пользователей социальной сети ВКонтакте. URL: <https://scm.etu.ru/assets/files/2025/sbornik/044-048.pdf> (дата обращения: 07.06.2025)

Анализ важности признаков

Таблица 11. Усредненная оценка важности признаков модели случайного леса

Код признака	Наименование признака	Важность (%)	Накоплено (%)
CT_1	Открытость–замкнутость	15.5	15.5
CT_7	Чувственность–твердость	15.5	31.0
SC_19	Гедонизм–индивидуальный приоритет	4.2	35.2
EY_1	Экстраверсия	4.0	39.2
CT_4	Беспечность–озабоченность	3.6	42.8
SC_3	Власть–нормативный идеал	3.4	46.2
LN_3	Циклотимность	3.3	49.5
BF_3	Самоконтроль–импульсивность	2.5	52.0

- Два наиболее важных признака (Кеттелла) — более 30% накопленной важности
- Первые восемь признаков — более 50% (всего 55 признаков)

Распределение значений С-индекса для предсказаний

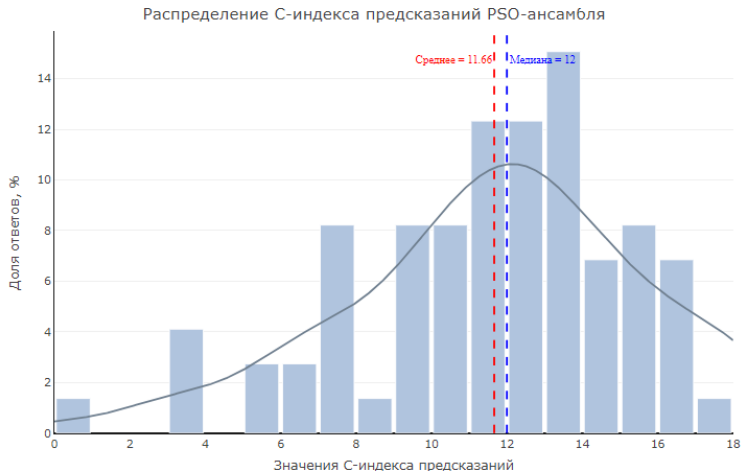


Рисунок 6. Распределение значений С-индекса для предсказаний PSO-ансамбля регрессоров

Обзор лучших моделей для каждого типа задач

Таблица 12. Обзор лучших моделей для каждого подхода и типа задач

Тип задач	Подход	Лучшая модель	C-индекс
Регрессия	ансамбль, mo	Рой частиц (Lasso-регрессия, пошаговая регрессия, CatBoost, ExtraTrees)	11.663
Классификация	ансамбль, ml	Рой частиц (kNN, SVM, логистическая Lasso-регрессия, XGBoost, LightGBM и др.)	11.625
Регрессия	ансамбль, chain	Рой частиц	11.613
Классификация	ансамбль, lp	Рой частиц / поиск по сетке	11.525
Классификация	ансамбль, mc	Генетический алгоритм / Рой частиц	11.263
Регрессия	multioutput	Lasso-регрессия	11.175
Регрессия	chained	Ridge-регрессия	11.062
Классификация	multilabel	k-ближайших соседей (kNN)	10.838
Ранжирование	списочное	MLP с ListNet@3	10.788
Классификация	multiclass	Логистическая Lasso-регрессия	10.663

Обозначения: mo — multioutput, ml — multilabel, lp — label powerset, mc — multiclass,

MLP — многослойный перцептрон, SVM — метод опорных векторов, kNN — метод k-ближайших соседей

Метрики качества RMSE и NDCG

- Корень из среднеквадратичной ошибки (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

где n — число объектов, y_i и \hat{y}_i — истинное и предсказанное значения

- Нормализованный дисконтированный совокупный прирост (NDCG@K):

$$\text{DCG@K} = \sum_{i=1}^K \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}, \quad \text{IDCG@K} = \sum_{i=1}^K \frac{2^{\text{rel}_i^*} - 1}{\log_2(i+1)}, \quad \text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}},$$

где K — глубина ранжирования, rel_i и rel_i^* — релевантность i -го элемента в ранжированном списке и в идеальном ранжировании

Сравнение регрессионных моделей (RMSE)

Таблица 13. Сравнение базовых регрессионных моделей по RMSE

Имя	Multioutput		Chained	
	без PCA	PCA	без PCA	PCA
Регрессия Lasso (L1)	2.018	2.036	2.018	2.030
Регрессия Ridge (L2)	2.025	2.037	2.028	2.044
Пошаговая регрессия	2.094	2.027	2.094	2.027
CatBoost	2.044	2.096	2.044	2.096
Случайный лес	2.069	2.131	2.070	2.133
LightGBM	2.074	2.128	2.074	2.128
Метод опорных векторов (SVR)	2.100	2.101	2.100	2.101
ExtraTrees	2.100	2.150	2.112	2.152
k-ближайших соседей (kNN)	2.162	2.151	2.162	2.151
Базовая константная	2.308	2.308	2.308	2.308
XGBoost	2.317	2.314	2.317	2.314
TabPFN	2.056			
MLP (BN, DropOut, регуляризация)	2.143			
MLP	2.442			

Результаты ансамбля регрессионных моделей (RMSE)

Таблица 14. Сравнение методов подбора весов ансамбля регрессионных моделей по RMSE

Метод подбора весов	Multioutput		Chained	
	все модели	топ-5	все модели	топ-5
Равные веса всех моделей	2.052	2.038	2.052	2.032
Вектор Шэпли (Shap)	2.047	2.035	2.046	2.033
Частичный перебор по сетке	2.052	2.026	2.045	2.035
Квадратичная оптимизация (QP)	2.109	2.093	2.111	2.070
Генетический алгоритм (GA)	2.035	2.036	2.044	2.032
Метод роя частиц (PSO)	2.049	2.031	2.065	2.026
Координатный спуск	2.097	2.048	2.089	2.040
Лин. регрессии с регуляризацией L1, L2, LightGBM, CatBoost, RF	Линейная регрессия		2.111	
	Линейная регрессия		2.091	

Обозначения:

топ-5 — подбор весов только для топ-5 моделей согласно метрике,

L1 и L2 — Lasso- и Ridge-модели регрессии, RF — случайный лес

Методы восстановления данных (RMSE)

Таблица 15. Восстановление значений незаполненных психометрических тестов с помощью базовых регрессионных моделей (RMSE)

Модель-регрессор	MICE	Soft Impute	Маски
Регрессия Lasso (L1)	2.059	2.046	2.098
CatBoost	2.054	2.092	2.105
Пошаговая регрессия	2.118	2.059	2.154
Линейная регрессия (OLS)	2.126	2.065	2.174
Регрессия Ridge (L2)	2.068	2.125	2.101
Случайный лес	2.081	2.085	2.113
ExtraTrees	2.083	2.126	2.133
LightGBM	2.083	2.110	2.136
Метод опорных векторов (SVR)	2.127	2.154	2.105
k-ближайших соседей (kNN)	2.107	2.150	2.142
Базовая константная	2.203	2.247	2.244
XGBoost	2.212	2.261	2.238

Ансамблевые методы восстановления для Soft Impute (RMSE)

Таблица 16. Весовые коэффициенты моделей и RMSE при разных методах подбора весов для ансамблей на восстановленных данных

Подбор весов	Веса моделей					RMSE
	Lasso L1	Пошагов.	LightGBM	Случ. лес	kNN	
Метод роя частиц (PSO)	0.001	0.481	0.038	0.475	0.005	2.038
Частичный перебор по сетке	0.000	0.500	0.000	0.500	0.000	2.043
Генетический алгоритм (GA)	0.281	0.369	0.109	0.189	0.052	2.044
Координатный спуск	0.019	0.422	0.067	0.305	0.187	2.052
Вектор Шэпли (Shar)	0.247	0.185	0.206	0.179	0.183	2.063
Равные веса всех моделей	0.200	0.200	0.200	0.200	0.200	2.064
Квадратичная оптимиз. (QP)	0.050	0.000	0.390	0.007	0.553	2.114

Обозначения:

Пошагов. — пошаговая регрессия,

Случ. лес — случайный лес (Random Forest),

kNN — метод k -ближайших соседей