Санкт-Петербургский государственный университет

Направление *02.04.03 «Математическое обеспечение и администрирование информационных систем»*

Кафедра информатики

***ГЛУШКОВ Егор Александрович***

Группа 23.М04-мм

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

Отчёт о прохождении производственной практики (преддипломной)

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

Консультант:
старший преподаватель кафедры информатики, к. т. н., Столярова В. Ф.

Санкт-Петербург
2025

# Оглавление

# Введение

Многие аспекты успешности человека обусловлены корректным определением карьерного пути, соответствующего его предпочтениям [5]. Карьерному самоопределению уделяются значительные ресурсы, в том числе со стороны государства. Золотым стандартом в сфере профориентации является глубинное интервью с экспертом, который поможет выявить сильные и слабые стороны личности. Однако этот подход является ресурсозатратным, и потому в качестве альтернативы развиваются дистанционные способы карьерного консультирования [14, 1]. Таким образом, актуальной является задача разработки моделей и алгоритмов выявления профессиональных предпочтений по доступным в дистанционном формате данным, таким как цифровые следы, профориентационные тесты [1, 6, 2].

Одним из инструментов для определения профессиональных интересов является модель RIASEC [7], которая предполагает наличие шести типов социально-профессиональной направленности личности: реалистический, исследовательский, артистический, социальный, предприимчивый и конвенциональный. Существует множество вариаций тестов, которые отражают показатели этой шкалы [20], 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 [20, 8].

Возникает актуальная задача определения кода Голланда по альтернативным данным. В настоящее время проводятся исследования взаимосвязи между кодом Голланда и результатами «Большой Пятерки» [11, 17, 10, 13] и цифровыми следами пользователей [6]. Несмотря на наличие работ, в которых предсказывается результат одного психометрического теста на основе другого теста или на основе некоторых признаков личности (например, комментариев, постов и фото пользователей социальной сети), до сих пор нет инструментов, позволяющих по результатам одного или комбинации сразу нескольких популярных пси-

хометрических тестов («Большая Пятерка», Кеттелла, Айзенка, Леонгарда, Шварца) предсказать код Голланда. В результате, пользователь мог бы получить информацию о своих профориентационных предпочтениях без прохождения теста Голланда. Кроме того, даже при прохождении последнего подобный инструмент мог бы уточнять результаты теста Голланда, проверять его непротиворечивость в соответствии с результатами других тестов.

# 1. Постановка задачи

Целью работы является разработка инструмента для автоматизации профориентации на основе определения кода Голланда по результатам психометрических тестов личности с использованием методов машинного обучения.

Для выполнения цели были поставлены следующие задачи:

1. Изучить и систематизировать существующие подходы к определению кода Голланда на основе психометрических данных.
2. Реализовать и сравнить методы машинного обучения для предсказания кодов Голланда.
3. Разработать алгоритмы восстановления результатов психометрических тестов в случае неполноты данных.
4. Создать прототип инструмента для определения профориентационных предпочтений.

# 2. Обзор

В работе представлен обзор модели Голланда и приведено краткое описание психометрических тестов личности. Проведен обзор предметной области: рассмотрены и проанализированы статьи, предлагающие различные подходы для решения задачи предсказания факторов психометрических тестов и для нахождения взаимосвязей между факторами.

## 2.1. Модель Голланда и психометрические тесты личности

Одним из основных инструментов оценки профессиональных интересов человека служит модель Голланда, также известная как модель RIASEC. Данная методика была разработана Джоном Льюисом Голландом (англ. J. L. Holland) в конце 1950-х годов, после чего им неоднократно дорабатывалась и развивалась [7]. В своей первой статье "Теория профессионального выбора" 1959 года американский исследователь сопоставляет различным типам личности профессиональные роды деятельности. Согласно Голланду, личности выбирают и преуспевают в той профессиональной среде, которая подходит их характеру, является отражением их базовых черт, при этом профессиональная карьерная среда классифицируется по тем типам личностей, которые в этой среде успешны. Таким образом, для определения профессиональных предпочтений достаточно определить социально-профессиональный тип личности.

В своих более поздних работах ученый выделяет следующие шесть типов личностей:

- реалистический (*Realistic*, R);
- исследовательский (*Investigative*, I);
- артистический (*Artistic*, A);
- социальный (*Social*, S);

- предприимчивый (*Enterprising*, E);
- традиционный (*Conventional*, C).

При этом определяется не единственный тип личности: оценивается принадлежность человека к каждому из типов, которые затем выстраиваются в порядке убывания их выраженности. Результат записывается в виде кода по первым буквам типов, что и дало название модели. Модель Голланда также можно представить как правильный шестиугольник с кодами в вершинах.

Для сравнения кода типа личности и кода его профессиональной среды Голланд вводит понятие конгруэнтности (согласованности). Среди мер конгруэнтности можно выделить C-индекс для трёхбуквенных кодов («верхних триад»):

$$C = 3\,(X_1, Y_1) + 2\,(X_2, Y_2) + 1\,(X_3, Y_3),$$

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

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

Помимо модели Голланда есть и другие способы определения профориентационных предпочтений. Одним из таких способов являются психологические тесты. Их основная цель — отразить некоторые черты личности человека в удобном числовом формате. Они помогают выявить ключевые черты характера, темперамент, ценности и поведенческие особенности.

Среди психологических тестов личности можно выделить следующие (в скобках указано количество факторов):

1. Опросник Леонгарда-Шмишека (10).

2. Личностный опросник Айзенка (4).

3. 16-факторный опросник Кеттелла (16).

4. Пятифакторный опросник личности («Большая пятерка»; 5).

5. Ценностный опросник Шварца (20).

## 2.2. Обзор предметной области

Применению методов математического моделирования в психологии в целом и нахождению взаимосвязей результатов различных психологических тестов между собой в частности посвящено множество научных работ.

В наши дни всё чаще встречается применение методов машинного обучения в психологии. Задачи могут быть различны.

- Предсказание кода Голланда на основе социально-демографических признаков [4]. 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 classificaion), многоцелевая классификация (multioutput classification). Авторы отмечают: в случае последовательного предсказания для многоцелевой регрессии порядок предсказания выходов важен. В качестве метрики авторы используют меру конгруэнтности — C -индекс. В качестве будущих работ предлагается обратить внимание на гексагональную меру, описанную в работе [15], и на подход с кластеризацией всего множества кодов (720 комбинаций) и дальнейшей классификацией на уже выделенные кластеры. Отметим, что лучшие результаты показал градиентный бустинг со значениями C-индекса 10.95 при решении задачи регрессии и 11.08 при решении задачи классификации.

- Оценка профессионального выбора (профессиональной рабочей среды среди трудоустроенных и профессиональных стремлений среди безработных) [9]. Пользователю по результатам прохождения теста Голланда предъявлялся список профессий, которым прежде уже был сопоставлен свой код Голланда; требовалось найти наиболее подходящие профессии. Наилучших результатов удалось достичь с помощью комбинации традиционных методов и ансамбля методов машинного обучения. В качестве традиционных методов использовалось сравнение значений мер конгруэнтности (простое совпадение главного фактора кода Голланда, оценка профилей — числовых значений кода Голланда — с помощью таких метрик, как коэффициент корреляции Пирсона и Евклидово расстояние). В ансамбль методов машинного обучения вошли следующие модели: многослойный перцептрон (нейронная сеть), метод k-ближайших соседей, регуляризованная регрессия, случайный лес.

- В статье [16] применяется логистическая регрессия для предсказания (классификации), какой путь выберут учащиеся: академический или профессиональный; предикторами служили значения факторов тестов Голланда и «Большой пятерки».

- Расширение списка профессий, поставленных в соответствие кодам Голланда, путем создания платформы для автоматизации профилирования вакансий [18]. Стоит отметить, что предсказание кодов Голланда решается как задача ранжирования с метрикой NDCG (англ. Normalized Discounted Cumulative Gain).

- Предсказание значений шкал теста «Большой пятерки» пользователей социальных сетей по следующим признакам: их посты, комментарии, репосты и численные характеристики аккаунта пользователя. Решалась задача бинарной классификации (шкалы теста были представлены бинарными путем сравнения с пороговым значением) с использованием моделей случайного леса и метода опорных векторов [12]. Подобная задача в работе [3] решалась

с помощью многослойного перцептрона. По схожим признакам (в т. ч. по указанной в профиле пользователя информации) для оценки темперамента (тест Айзенка EPQ) в статье [2] использовались модели CatBoost и случайный лес. В работе [19] предсказание результатов тестов «Большой пятерки», Шварца и других по извлекаемым из профилей в социальной сети численным признакам (число друзей, постов, подписок, длина поля с личным описанием, длина постов и др.) осуществляется с помощью модели XGBoost (eXtreme Gradient Boosting).

## 2.3. Выводы

Для предсказания кода Голланда могут быть использованы различные идеи, приведенные в данном обзоре, например, представление задачи как регрессии/классификации с множественными выходами (multioutput), классификации с несколькими метками; для многоцелевой регрессии могут быть использованы различные метрики: не только усредненные MSE или RMSE, часто применяемые в таких задачах, но и меры конгруэнтности (C-индекс), косинусное расстояние, коэффициент корреляции; приведены различные методы машинного обучения, в том числе и те, с помощью которых были достигнуты наилучшие результаты: в первую очередь, это градиентный бустинг (CatBoost, XGBoost) и случайный лес.

Несмотря на разнообразие работ, наличие среди них тех, где по результатам одних психометрических тестов предсказываются другие, до сих пор нет инструментов, позволяющих по результатам сразу нескольких популярных психометрических тестов — «Большой пятерки», Кеттелла, Айзенка, Леонгарда, Шварца — предсказать код Голланда. Таким образом, пользователь мог бы получить информацию о своих профориентационных предпочтениях без прохождения теста Голланда. Кроме того, даже при прохождении последнего подобный инструмент мог бы уточнять результаты теста Голланда, проверять его непротиворечивость в соответствии с результатами других тестов.

# 3. Предсказание кода Голланда

В данном разделе описываются различные подходы к предсказанию кода Голланда на полных данных (без пропусков) и на данных с пропусками, требующими восстановления.

## 3.1. Предсказание кода Голланда на полных данных

Данные по психометрическим тестам представляют собой набор признаков, принимающих целочисленные значения. Пример данных психометрических тестов приведен в таблице 1. Предсказание кода Голланда представляет собой предсказание шести чисел, отражающих степень выраженности типов личности (кодов Голланда). В качестве альтернативы предсказанием кода может служить ответ как в виде одного буквенного значения, так и набора из трех значений, «верхней триады» — тройки наиболее выраженных факторов. В случае, если в таком наборе задан порядок, то речь может идти о предсказании рангов кодов (от менее выраженного к наиболее выраженному).

Таблица 1. Пример данных психометрических тестов

| | Большая Пятёрка | | | ... | Леонгард | | Голланд | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| id | BF1 | BF2 | BF3 | ... | LN9 | LN10 | HL1 | HL2 | HL3 | HL4 | HL5 | HL6 |
| 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 |
| 5 | 48 | 42 | 53 | ... | 12 | 6 | 6 | 7 | 8 | 7 | 10 | 4 |

Таким образом, задача об определении кода Голланда по результатам психометрических тестов личности может быть сведена к следующим задачам:

1. Регрессия со множественными выходами (многоцелевая регрессия, англ. *multioutput/multitarget*);

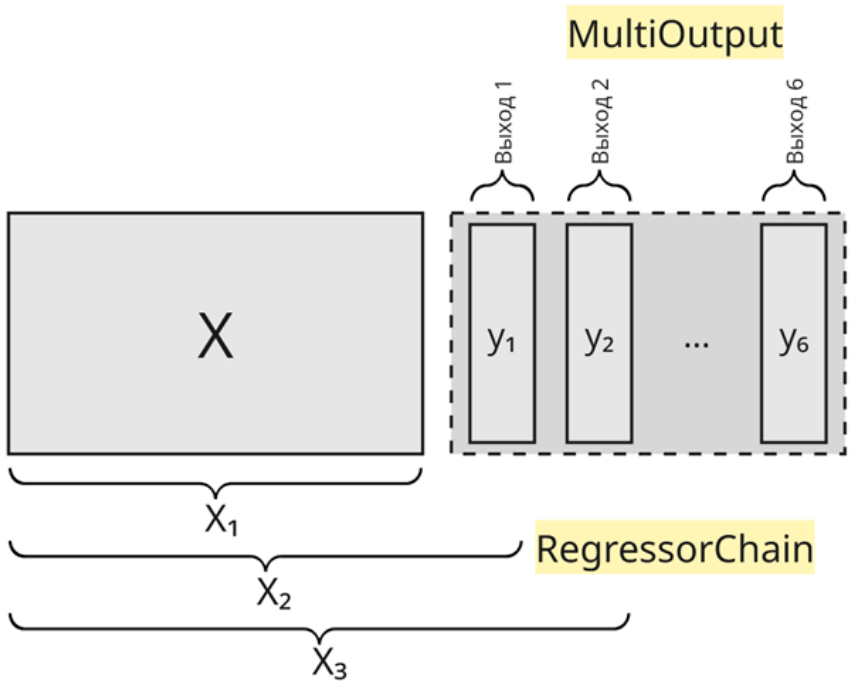


Рисунок 1. Регрессия со множественными выходами

2. Классификация;
3. Ранжирование;
4. Ансамблевые модели.

### 3.1.1. Регрессия

Предсказание сразу нескольких целевых числовых переменных представляет собой регрессию со множественными выходами. Существует три основных подхода к решению данной задачи:

- Использование моделей, поддерживающих многозадачность «из коробки» (линейная регрессия, метод k-ближайших соседей);
- Независимые предсказания каждого из выходов (multioutput);
- Предсказания выходов по цепочке, когда последний предсказанный выход становится частью признакового пространства для предсказания следующего выхода (см. рисунок 1).

Оцениваются результаты решения задачи регрессии со множественными выходами с помощью следующих усредненных по ответам метрик на тестовой выборке:

- усредненная среднеквадратичная ошибка (RMSE);
- усредненный C-индекс (см. 2.1).

Лучшее качество обеспечивается при минимальных значениях RMSE и при максимальных значениях C-индекса (так как является мерой согласованности). Чем больше C-индекс, тем большее сходство имеют два сравниваемых между собой профиля. Использование C-индекса позволяет сравнивать результаты не только с другими регрессионными задачами, но и, например, с задачами классификациями. Стоит отметить, что в процессе обучения некоторые модели также оценивают важность признаков: ансамблевые модели на основе решающих деревьев (например, случайный лес, градиентные бустинги) и линейные регрессоры.

Для нахождения взаимосвязей между факторами модели Голланда и другими психометрическими тестами были использованы следующие регрессионные модели в качестве базовых:

- модели на основе линейной регрессии (линейная регрессия, L1- и L2-регуляризованная регрессия, пошаговая регрессия);
- модели на основе ансамблей деревьев с бутстрэпингом (случайный лес, ExtraTrees);
- модели градиентного бустинга (XGBoost, LightGBM, CatBoost);
- иные «классические» модели (kNN, SVR);
- нейросетевой подход (MLP, foundation-модель TabPFN);
- базовая константная модель для сравнительной проверки результатов других моделей.

При отсутствии пропусков (т. е. при работе с полными данными) может быть целесообразным уменьшение размерности с помощью метода главных компонент (англ. Principal Component Analysis). Некоторые факторы различных тестов могут оценивать одни и те же аспекты личности, среди связей могут наблюдаться корреляции. Уменьшение размерности позволяет «схлопывать» похожие или одинаковые факторы тестов.

### 3.1.2. Классификация

Решение задачи классификации кода Голланда может быть сведено к следующим подходам:

- Многоклассовая (multiclass) классификация: обучается один классификатор на шесть классов. В качестве ответа берутся те три кода, которые имеют наибольшие вероятности.

- Многометочная (multilabel) классификация: обучаются шесть бинарных классификаторов, позволяющих дать ответ на вопрос, входит ли соответствующий код в «верхнюю триаду». Для каждого из шести кодов получается вероятность вхождения в верхнюю тройку, откуда в соответствии с наибольшей предсказанной вероятностью отбирается тройка кодов.

- Классификация Label Powerset. Всего существует 20 различных трехбуквенных комбинаций кода Голланда, в которых не учитывается порядок кодов. На выходе — тройка кодов. По причине отсутствия возможности учесть порядок кодов, для данного подхода не представляется возможным использовать C-индекс, оценка которого основывается именно на порядке кодов.

По аналогии с решением задачи регрессии также может быть использован метод главных компонент. Для оценки решения задачи классификации была использована метрика точности Top-K (Top-K accuracy), которая измеряет долю случаев, когда правильный класс оказался среди

трёх наиболее вероятных (по мнению модели) предсказанных классов. Для сравнения с результатами других подходов используется C-индекс.

Для задачи классификации были взяты те же модели, что и для регрессии, за исключением нейросетевых моделей, а также с добавлением логистических регрессий вместо обычных линейных регрессий и классификатора Наивного Байеса.

### 3.1.3. Ранжирование

Представим задачу как списочное ранжирование. В отличие от поточечного ранжирования, где каждый элемент оценивается по шкале релевантности (что фактически сводится к задаче регрессии), или попарного ранжирования, в котором моделируется относительный порядок пар элементов (бинарная классификация), listwise-ранжирование рассматривает весь список результатов как единый объект.

Списковая постановка требует двух ключевых компонентов: во-первых, определения скоринговой функции, во-вторых, выбора функции потерь, оптимизирующей качество выдачи списка. В качестве скоринговой функции обычно используют многослойный перцептрон или более сложные архитектуры — глубокую перекрёстную сеть (Deep & Cross Network) либо трансформер с self-attention.

Типичными функциями потерь для списочного ранжирования являются Normalized Discounted Cumulative Gain (NDCG) и его дифференцируемые аппроксимации (ApproxNDCG, LambdaRank), а также специальные дифференцируемые методы, например ListNet, который минимизирует кросс-энтропию распределений релевантности и оптимизирует вероятность корректного попадания в топ-k (в нашем случае top-1 и top-3).

Для оценки качества обученных моделей в качестве метрики часто используют NDCG@3, отражающую «полезность» первых трёх элементов выдачи с учётом их позиций и релевантности. Списочный подход обычно обеспечивает более высокую точность, однако требует значительных вычислительных ресурсов. В задачах с небольшим размером списка (шесть элементов) и ограниченным объёмом данных это ограни-

чение, как правило, не является критическим.

### 3.1.4. Ансамблевые модели

Для улучшения качества прогноза может быть использовано взвешенное ансамблирование моделей (линейный блендинг), когда итоговый ответ вычислялся как линейная комбинация предсказаний различных моделей с оптимизированными коэффициентами. Другим подходом является обучение мета-модели на предсказаниях базовых моделей — стекинг.

Для стекинга были взяты следующие базовые модели:

- линейные регрессии с различными параметрами регуляризации;
- L1- и L2-регуляризованные регрессии, LightGBM, CatBoost, Случайный лес.

В обоих случаях метамоделью выступает обычная линейная регрессия. Важно отметить, что в первом случае при отсутствии регуляризации получалась бы линейная комбинация линейных моделей, что также является линейной моделью, и именно по этой причине добавляется нелинейная составляющая. Обучение базовых моделей происходит на обучающей выборке, обучение метамодели — на валидационной выборке, оценка метрик — на тестовой выборке.

Для линейного блендинга требуется найти, с каким весом будет входить в итоговое предсказание каждое из предсказаний базовых моделей. Таким образом, линейная комбинация весов моделей на их предсказание и будет итоговым предсказанием. Подбор весов, как и подгонка модели под данные, происходит на валидационной выборке. Применяются следующие подходы для подбора весов:

– Равные веса всех моделей
  каждой базовой модели присваивается одинаковый вклад, что упрощает ансамблирование и служит «базовой линией»;

– Вектор Шэпли
  распределение общего вклада каждого элемента ансамбля на основе их маргинального вклада;

– Частичный перебор по сетке
  поиск решений на предварительно заданной сетке возможных значений параметров (весов), при увеличении числа элементов для поиска вклада каждого предполагается использование подвыборки заданной сетки;

– Квадратичная оптимизация (QP)
  решение задачи минимизации взвешенной суммы ошибок ансамбля как задачи квадратичного программирования с ограничениями;

– Генетический алгоритм (GA)
  эволюционный поиск с выбором лучших представителей популяций (комбинаций весов), их скрещиванием и мутациями;

– Метод роя частиц (PSO)
  оптимизация весов с помощью популяции частиц, которые перемещаются по пространству решений с соответствии с комбинацией собственного и глобального (популяции) оптимального пути;

– Координатный спуск
  итеративная оптимизация веса каждой модели при фиксации остальных.

Стоит отметить, что лишь квадратичная оптимизация требует наличия градиента функции потерь.

## 3.2. Восстановление данных отсутствующих психометрических тестов

В исходной задаче предполагается наличие пропусков (неполноты) в данных, поскольку пользователь мог и не успеть пройти пять пси-

хометрических тестов перед тем, как запрашивается предсказание его кодов Голланда. Существуют следующие подходы для восстановления данных отсутствующих тестов:

1. MICE
   множественная импутация цепочными уравнениями, при которой для каждого признака поочерёдно строятся регрессионные модели на основе остальных и заполняются пропуски в нескольких итерациях;

2. Matrix Soft Impute
   метод восстановления матрицы с помощью мягкого порогового SVD-разложения, приводящего к низкоранговой аппроксимации;

3. Применение масок для пропусков
   добавление бинарных индикаторов отсутствия данных в качестве признаков, чтобы модель учитывала факт пропуска напрямую;

4. Взвешенное ансамблирование (блендинг) по той комбинации тестов, которые были заполнены пользователем; всего 31 комбинация наличия тестов.

Стоит отметить, что при обработке одной записи на вход подход MICE требует поступления и других данных тоже, как бы обучаясь каждый раз даже во время предсказания, также возможны проблемы со сходимостью при скоррелированных признаках. Взвешенное ансамблирование путем нахождения весов для 31 комбинации наличия тестов может быть достаточно трудозатратной в вычислительном плане задачей. Применение масок не восстанавливает значения пропущенных факторов: лишь использует знание об их пропусках. Применение сингулярного разложения может быть затратным для больших матриц.

# 4. Реализация подходов для определения кода Голланда

## 4.1. Используемые данные

Данные для исследования были собраны с помощью опроса, размещенного в веб-приложении на базе платформы VK Mini Apps «Психологические тесты»[1]. Приложение находится в открытом доступе и позволяет пользователям проходить различные психометрические опросы. 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»).

Всего имеются данные по 1278 пользователям: у 339 есть данные по всем тестам, у 939 пользователей отсутствует один или два теста. Пример части преобразованного набора данных представлен в таблице 1. Диаграмма размаха для факторов модели Голланда представлена на рисунке 2. Матрица корреляций кодов Голланда — рисунок 3. Описательная статистика по всем факторам всех психометрических тестов приведена в Приложении А в таблице 14.

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.

## 4.2. Определение кода Голланда. Сравнение моделей

Реализация подходов по определению профориентационных предпочтений на основе психометрических тестов личности производилась

[1]Мини-приложение «Психологические тесты» (платформа «VK Mini Apps»). URL: https://vk.com/app7794698 (дата обращения: 17.05.2025).

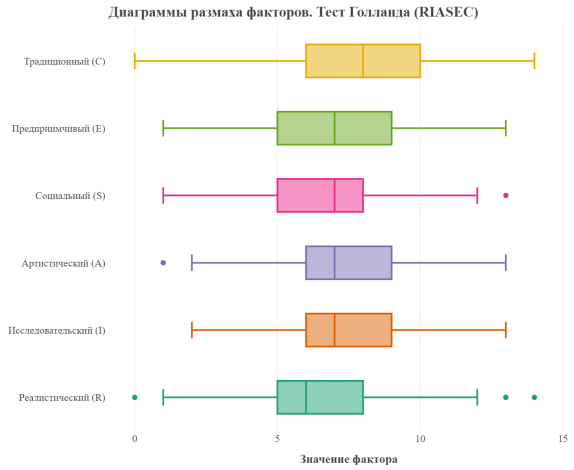


Рисунок 2. Диаграмма размаха факторов модели Голланда

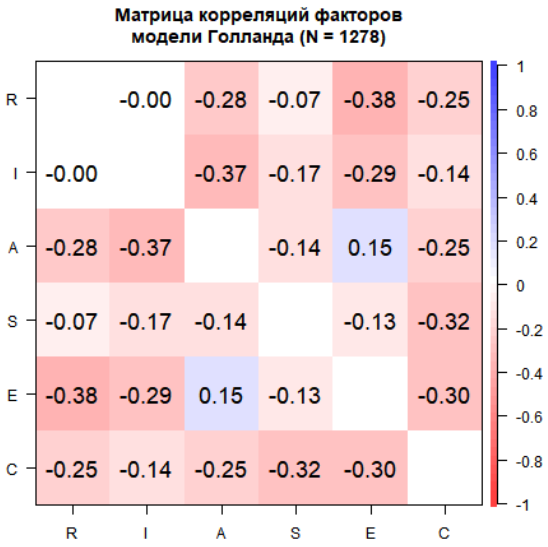


Рисунок 3. Матрица корреляций кодов Голланда

на языке R, являющимся открытым и свободным программным обеспечением, преимущественно с использованием библиотек *data.table* и *R6Class*, позволяющим эффективно оперировать данными. Нейросетевые модели были реализованы с помощью *Python* и фреймворка *PyTorch*, также открытых и свободных ПО. Исходный код используемых моделей и поискового анализа представлен в репозитории[2].

Перед проведением сравнительного анализа данные были получены и преобразованы из json-файлов, проведена валидация данных (допустимые значения по каждому из факторов и иные ограничения), проведена стандартизация (нормализация). Для некоторых моделей (если не указано иное) итоговое значение метрики — это лучшее из значений модели на полном наборе данных и на наборе после уменьшения размерности.

Для обеспечения сравнимости различных подходов (регрессия, классификация, ранжирование) между собой основное внимание было уделено метрике C-индекс. Такой выбор позволяет сравнивать результаты и с другими научными работами. Так, на основе социо-

[2]GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности. URL: https://github.com/ExP98/Diploma_Holland (дата обращения: 17.05.2025).

Таблица 2. Сравнение регрессионных моделей по метрике C-индекс

| **Модель** | **MO** | **MO PCA** | **Chain** | **Chain 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 | 10.600 | 10.900 | 10.600 | 10.900 |
| CatBoost | 10.688 | 10.812 | 10.688 | 10.812 |
| Random Forest | 10.625 | 10.475 | 10.812 | 10.588 |
| Линейная регрессия (OLS) | 10.688 | 10.800 | 10.688 | 10.800 |
| LightGBM | 10.750 | 10.425 | 10.750 | 10.425 |
| kNN | 10.525 | 10.400 | 10.525 | 10.400 |
| XGBoost | 9.162 | 9.725 | 9.162 | 9.725 |
| Constant baseline | 9.000 | 9.000 | 9.000 | 9.000 |
| TabPFN | 10.562 | | | |
| MLP (BatchNorm, DropOut, регуляризация) | 10.462 | | | |
| MLP | 10.275 | | | |

** Обозначения:*
*MO — Multioutput, предсказание выходных переменных независимо друг от друга,*
*Chain — Предсказание выходных переменных по цепочке,*
*PCA — метод главных компонент (уменьшение размерности).*

демографических данных в работе [4] достигается значение C-index = 10.95 для регрессии и C-index = 11.08 для классификации.

Результаты определения кода Голланда как задачи регрессии (метрика C-индекс) приведены в таблице 2. В таблице приведен константный предсказатель, C-индекс которого равен 9. В сравнении с ним модель XGBoost показывает низкие результаты. Наилучшие результаты показывает регуляризованная регрессия (Lasso и Ridge): C-index = 11.175 и C-index = 11.062. Высокий результат показывает модель Extremely Randomized Trees (ExtraTrees): C-index = 11.1. Лучшая из нейросетевых моделей, foundation-модель TabPFN, C-index = 10.562, показывает результат лишь лучше XGBoost и на одном уровне с ме-

Таблица 3. Сравнение моделей по метрике RMSE

| **Модель** | **MO** | **MO PCA** | **Chain** | **Chain PCA** |
|---|---|---|---|---|
| Регрессия Lasso (L1) | 2.018 | 2.036 | 2.018 | 2.030 |
| Линейная регрессия (OLS) | 2.155 | 2.019 | 2.155 | 2.019 |
| 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 |
| Random Forest | 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 |
| kNN | 2.162 | 2.151 | 2.162 | 2.151 |
| Constant baseline | 2.308 | 2.308 | 2.308 | 2.308 |
| XGBoost | 2.317 | 2.314 | 2.317 | 2.314 |
| TabPFN | 2.056 | | | |
| MLP (BatchNorm, DropOut, регул-я) | 2.143 | | | |
| MLP | 2.442 | | | |

** Обозначения:*
*MO — Multioutput, предсказание выходных переменных независимо друг от друга,*
*Chain — Предсказание выходных переменных по цепочке,*
*PCA — метод главных компонент (уменьшение размерности).*

тодом k-ближайших соседей. Стоит отметить, что результаты моделей при различных подходах, MultiOutput и Chain, практически идентичны. В то же время попарно для каждого из подходов лучшие результаты модели показывают на наборе данных меньшей размерности (после применения метода главных компонент, PCA), кроме регуляризованных линейных регрессий. Сравнение регрессионных моделей по метрике RMSE приведено в таблице 3.

На примере модели случайного леса (Random Forest) в таблице 4 приведен анализ важности признаков. Так, два фактора из модели Кеттелла покрывают более 30% важности всех 55 факторов, 8 факторов —

Таблица 4. Важность признаков модели Random Forest

| 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 – индив. приоритет | 4.2 | 35.2 |
| EY_1 | Экстраверсия | 4.0 | 39.2 |
| CT_4 | Беспечность – Озабоченность | 3.6 | 42.8 |
| SC_3 | Власть – нормат. идеал | 3.4 | 46.2 |
| LN_3 | Циклотимность | 3.3 | 49.5 |
| BF_3 | Самоконтроль – импульсивность | 2.5 | 52.0 |
| BF_4 | Эмоц. устойчивость – неустойчивость | 2.5 | 54.5 |

более 50%. В Random Forest важность признака (gain) — это усреднённая по всем деревьям сумма уменьшений критерия нечистоты (Gini или энтропии) на узлах, где при разбиении использовался этот признак. При аналогичном анализе важности признаков с помощью моделей градиентного бустинга получаются схожие результаты: наиболее важными признаются схожие признаки, но они имеют меньшую важность.

Сравнение методов подбора весов ансамбля регрессионных моделей представлено в таблице 5. Лучшим методом подбора весов для моделей линейного блендинга (взвешенного ансамблирования) является метод роя частиц (PSO), C-index = 11.663. В таблице 6 приведены веса входящих в лучшую PSO-модель базовых регрессоров: наибольший вклад вносит Lasso-регрессор (43.2%), а также пошаговая регрессия. Модели стекинга показывают результаты хуже, чем модели линейного блендинга.

На рисунке 4 показана гистограмма распределения значений C-индекса для предсказаний PSO-ансамбля.

Сравнение моделей базовых классификаторов в разрезе трех главных подходов для классификации показано в таблице 7. Сравнение про-

Таблица 5. Сравнение методов подбора весов ансамбля регрессионных моделей

| Метод подбора весов | MO | MO избр. | Chain | Chain избр. |
|---|---|---|---|---|
| Равные веса всех моделей | 11.063 | 11.088 | 11.050 | 11.013 |
| Вектор Шэпли (Shap) | 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 |
| Линейные регрессии с регуляризацией | Линейная регр. | | 10.887 | |
| Lasso, Ridge, LightGBM, CatBoost, RF | Линейная регр. | | 10.688 | |

** MO — Multioutput, избр. — подбор весов только для топ-5 моделей согласно C-индексу*

Таблица 6. Весовые коэффициенты моделей и C-индекс

| Метод подбора весов | Веса моделей | | | | C-индекс |
|---|---|---|---|---|---|
| | Lasso L1 | Пошаговая регр. | CatBoost | ExtraTrees | |
| PSO | 0.432 | 0.327 | 0.150 | 0.091 | 11.663 |

изводилось по метрике Top-K точность. Подходы multiclass и multilabel показывают схожие результаты и опережают подход label powerset по метрикам точности Top-1 и Top-5. Можно заметить, что большинство моделей в 98%–100% случаев предсказывают в тройке кодов хотя бы один, который действительно есть в фактических данных; в 70% и более угадываются хотя бы два кода. Лишь примерно в 15% правильно предсказываются все три кода.

Таблица 8 — Сравнение лучших моделей классификации. На первом месте с C-index = 10.838 метод k-ближайших соседей (подход multilabel), за ним логистическая Lasso-регрессия (multiclass) с C-index = 10.663.

Сравнение методов подбора весов ансамбля классификаторов представлено в таблице 9. Лучшим методом подбора весов для моделей линейного блендинга (взвешенного ансамблирования) является метод роя

Таблица 7. Сравнение подходов к классификации (Top-K accuracy)

| **Модель** | **Multiclass** | | | **Multilabel** | | | **Label Powerset** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Top1 | Top2 | Top3 | Top1 | Top2 | Top3 | Top1 | Top2 | Top3 |
| kNN | 0.988 | 0.713 | 0.125 | 1.000 | 0.763 | 0.113 | 0.975 | 0.650 | 0.175 |
| Логист. L1-регр. | 1.000 | 0.700 | 0.163 | 1.000 | 0.700 | 0.163 | 0.988 | 0.638 | 0.100 |
| XGBoost | 1.000 | 0.700 | 0.113 | 0.975 | 0.675 | 0.100 | 0.963 | 0.625 | 0.113 |
| Логист. L2-регр. | 1.000 | 0.700 | 0.150 | 0.988 | 0.700 | 0.213 | 0.988 | 0.675 | 0.088 |
| Наивный Байес | 0.975 | 0.700 | 0.150 | 0.988 | 0.700 | 0.150 | 0.988 | 0.688 | 0.163 |
| ExtraTrees | 0.996 | 0.725 | 0.150 | 1.000 | 0.775 | 0.146 | 0.975 | 0.688 | 0.200 |
| SVM | 1.000 | 0.742 | 0.146 | 1.000 | 0.721 | 0.138 | 0.975 | 0.675 | 0.213 |
| Random Forest | 1.000 | 0.742 | 0.158 | 0.996 | 0.738 | 0.146 | 0.988 | 0.638 | 0.225 |
| CatBoost | 0.988 | 0.788 | 0.113 | 0.988 | 0.788 | 0.113 | 0.988 | 0.700 | 0.163 |
| LightGBM | 0.975 | 0.563 | 0.050 | 0.975 | 0.700 | 0.100 | 0.950 | 0.525 | 0.050 |

частиц (PSO), C-index = 11.625. В таблице 10 приведены веса регрессоров, веса которых определены с помощью PSO: наибольший вклад вносят kNN и SVM. Таким образом, результаты лучшего взвешенного ансамбля классификаторов лишь немного хуже лучшего регрессионного ансамбля.

Сравнение моделей ранжирования отражено в таблице 11. Лучшая из моделей, многослойный перцептрон с функцией потерь ListNet@3, показывает C-index = 10.788, что заведомо хуже, чем лучшие из базовых регрессоров и классификаторов.

Лучшие из моделей по каждому рассматриваемому типу задач приведены в таблице 12. Двумя лучшими моделями оказались линейные блендинги, веса которых подобраны методом роя частиц: это подходы multioutput для регрессии (C-index = 11.663) и multilabel для классификации (C-index = 11.625). Базовые модели показывают результаты хуже, чем их комбинация в виде взвешенных ансамблей.

Таблица 8. Сравнение классификаторов

| **Классификатор** | **Подход** | **C-индекс** | **Top1** | **Top2** | **Top3** |
|---|---|---|---|---|---|
| kNN | Multilabel | 10.838 | 1.000 | 0.763 | 0.113 |
| Логистич. регр. (L1) | Multiclass | 10.663 | 1.000 | 0.700 | 0.163 |
| XGBoost | Multiclass | 10.638 | 1.000 | 0.700 | 0.113 |
| Логистич. регр. (L2) | Multiclass | 10.500 | 1.000 | 0.700 | 0.150 |
| Наивный Байес | Multilabel | 10.350 | 0.988 | 0.700 | 0.150 |
| ExtraTrees | Multilabel | 10.013 | 1.000 | 0.775 | 0.146 |
| SVM | Multilabel | 9.875 | 1.000 | 0.721 | 0.138 |
| Random Forest | 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 |
| Baseline (случайный) | – | 9.000 | 0.950 | 0.500 | 0.050 |

## 4.3. Реализация восстановления данных психометрических тестов

Восстановление значений незаполненных психометрических тестов представлено в таблице 13. Хуже всего (согласно C-индексу) себя показывает подход множественной импутации. Лучшая модель для восстановления пропущенных значений тестов — это Lasso-регрессор и подход мягкого матричного восстановления данных (C-index = 10.09). Значения метрик на восстановленных данных заметно ниже, чем на полных данных для аналогичных моделей регрессии и классификации. Тем не менее, эти значения превосходят результаты моделей на данных, разделенных по психометрическим тестам.

## 4.4. Прототип инструмента для определения профориентационных предпочтений

Прототип инструмента для определения профориентационных предпочтений был реализован на R Shiny. На рисунке 5 представлен общий порядок расчетов разработанного программного модуля. В последовательности шагов есть обработка данных, восстановление пропусков с помощью модели мягкого матричного восстановления, предсказывает-


Рисунок 4. Распределение значений C-индекса для предсказаний PSO-ансамбля

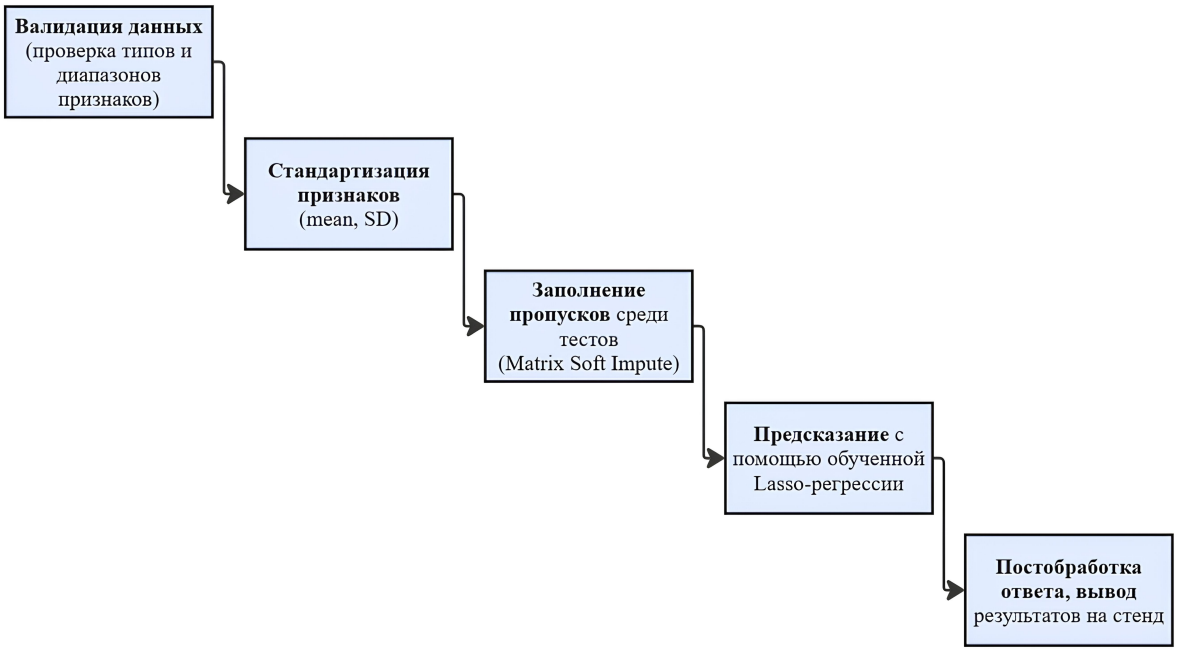


Рисунок 5. Итоговая последовательность вычислительных шагов

Таблица 9. 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 10. Весовые коэффициенты моделей и C-индекс

| Метод подбора | Веса моделей | | | | | C-индекс |
|---|---|---|---|---|---|---|
| весов | kNN | SVM | Logit L1 | XGBoost | LightGBM | |
| PSO | 0.291 | 0.164 | 0.191 | 0.183 | 0.151 | 11.625 |

ся с помощью Lasso-регрессии. Результат отображается на стенде 6 в виде последовательности кодов и текстового описания.

Таблица 11. Сравнение моделей ранжирования

| **Функция потерь** | **C-индекс** | | | **NDCG@3** | | |
|---|---|---|---|---|---|---|
| | Deep and Cross | Listwise Transformer | MLP | Deep and Cross | Listwise 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 |

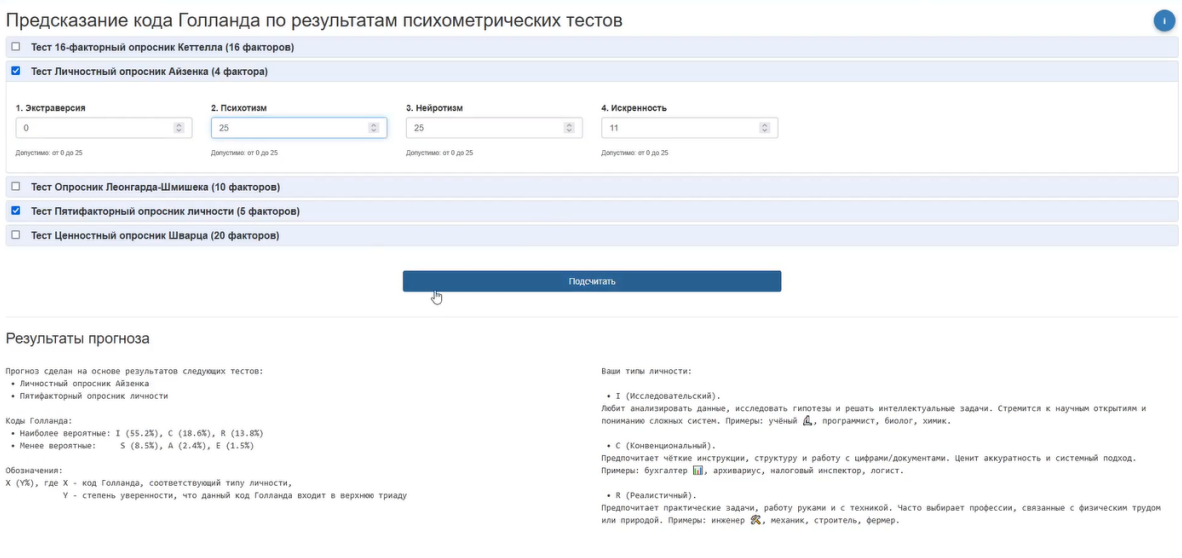


Рисунок 6. Интерфейс прототипа инструмента профориентации

# Заключение

Целью данной работы являлась разработка инструмента для автоматизации профориентации на основе определения кода Голланда по результатам психометрических тестов личности с использованием методов машинного обучения.

Для выполнения цели были решены следующие задачи:

1. Изучены существующие подходы к определению кода Голланда, поставлены задачи:
   - регрессии (multioutput, chain);
   - классификации (multiclass, multilabel, label powerset);
   - линейного блендинга базовых моделей;

- ранжирования.

2. Реализованы модели для предсказания кодов Голланда:
   лучшие результаты у моделей линейного блендинга с весами базовых моделей, подобранными на основе метода роя частиц:

   - Ансамбль multioutput-регрессоров: Lasso-регрессия, пошаговая регрессия, CatBoost, ExtraTrees (C-index = 11.663);
   - Ансамбль multilabel-классификаторов: kNN, SVM, логистическая Lasso-регрессия, XGBoost, LightGBM и др. (C-index = 11.625);
   - Показано, что классические методы машинного обучения превосходят нейросетевые в задаче предсказания кода Голланда.

3. Реализованы алгоритмы восстановления результатов психометрических тестов: наилучший результат достигнут методом мягкой импутации в сочетании с Lasso-регрессией (C-index = 10.09).

4. Создан прототип инструмента для определения профориентационных предпочтений: стенд на основе R Shiny.

Исходный код реализации стенда, представленных моделей и поискового анализа представлен в репозитории[3].

[3]GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности. URL: https://github.com/ExP98/Diploma_Holland (дата обращения: 17.05.2025).

# Список литературы


[1] Artificial Intelligence for Career Guidance – Current Requirements and Prospects for the Future / Stina Westman, Janne Kauttonen, Aarne Klemetti et al. // IAFOR Journal of Education. — 2021. — 08. — Vol. 9. — P. 43–62.

[2] Automating the Temperament Assessment of Online Social Network Users / Valerii Oliseenko, Anastasia Ivaschenko, A. Korepanova, T. Tulupyeva // Doklady Mathematics. — 2024. — 02. — Vol. 108.

[3] Başaran Seren, Ejimogu Obinna. A Neural Network Approach for Predicting Personality From Facebook Data. — 2021. — 07.

[4] Bogacheva Eugenia, Tatarenko Filipp, Smetannikov Ivan. Predicting Vocational Personality Type from Socio-demographic Features Using Machine Learning Methods. — 2020. — 10. — P. 93–98.

[5] Career Competencies and Career Success: On the Roles of Employability Activities and Academic Satisfaction During the School-to-Work Transition / Alessandro Lo Presti, Vincenza Capone, Ada Aversano, Jos Akkermans // Journal of Career Development. — 2021. — 02. — Vol. 49. — P. 089484532199253.

[6] Chekalev A., Khlobystova A., Abramov M. Community Theme Analyser: Predicting Career Guidance in Online Social Networks // Proceedings of the Eighth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'24), Volume 2 / Ed. by Sergey Kovalev, Igor Kotenko, Andrey Sukhanov et al. — Cham : Springer Nature Switzerland, 2024. — P. 153–162.

[7] Holland J.L. Making Vocational Choices: A Theory of Vocational Personalities and Work Environments. Prentice-Hall series in counseling and human development. — Prentice-Hall, 1985. — ISBN: 9780135475973. — URL: https://books.google.ru/books?id=8QxBAAAAMAAJ.

[8] Interested and employed? A national study of gender differences in basic interests and employment / Kevin Hoff, Kenneth Granillo-Velasquez, Alexis Hanna et al. // Journal of Vocational Behavior. — 2024. — 05. — Vol. 148. — P. 103942.

[9] Investigating machine learning's capacity to enhance the prediction of career choices / Q. Chelsea Song, Hyun Joo Shin, Chen Tang et al. // Personnel Psychology. — 2022. — 06. — Vol. 77. — P. n/a–n/a.

[10] M. Schuerger J. Career assessment and the sixteen personality factor questionnaire. — 1995.

[11] Mason Rod, Roodenburg John. Personality and vocational interest typologies associated with better coping and resilience in paramedicine: A review of two models // Paramedicine. — 2023. — 09. — Vol. 21.

[12] Personality Traits Prediction from Vkontakte Social Media / Maksim Stankevich, Nikolay Ignatiev, Ivan Smirnov, Natalia Kiselnikova // Voprosy kiberbezopasnosti. — 2019. — 01. — P. 80–87.

[13] Personality Traits Systematically Explain the Semantic Arrangement of Occupational Preferences / Jumpei Yamashita, Ritsuko Iwai, Haruo Oishi, Takatsune Kumada // Journal of Individual Differences. — 2024. — 08. — Vol. 45. — P. 201–217.

[14] Pordelan Nooshin, Hosseinian Simin. Design and development of the online career counselling: a tool for better career decision-making // Behaviour Information Technology. — 2020. — 07. — Vol. 41. — P. 1–21.

[15] Prediger Dale, Vansickle Timothy. Locating occupations on Holland's hexagon: Beyond RIASEC // Journal of Vocational Behavior. — 1992. — 04. — Vol. 40. — P. 111–128.

[16] RIASEC Interests and the Big Five Personality Traits Matter for Life Success—But Do They Already Matter for Educational Track

Choices? / Nele Usslepp, Nicolas Hübner, Gundula Stoll et al. // Journal of Personality. — 2020. — 03. — Vol. 88.

[17] Rúa Sandra M. Hurtado, Stead Graham B., Poklar Ashley E. Five-Factor Personality Traits and RIASEC Interest Types: A Multivariate Meta-Analysis // Journal of Career Assessment. — 2019. — Vol. 27, no. 3. — P. 527–543. — https://doi.org/10.1177/1069072718780447.

[18] Silva Amila, Lo Pei-Chi, Lim Ee-Peng. JPLink: On Linking Jobs to Vocational Interest Types. — 2020. — 05. — P. 220–232. — ISBN: 978-3-030-47435-5.

[19] Titov Sergey, Novikov Pavel, Mararitsa Larisa. Full-scale Personality Prediction on VKontakte Social Network and its Applications. — 2019. — 11. — P. 317–323.

[20] What Do Interest Inventories Measure? The Convergence and Content Validity of Four RIASEC Inventories. / Chu Chu, Mary Russell, Kevin Hoff et al. — 2022. — 05.

Таблица 12. Обзор лучших моделей для каждого типа задач

| **Тип задач** | **Подход** | **Лучшая модель** | **C-индекс** |
|---|---|---|---|
| Регрессия | Блендинг, mo | PSO (Lasso-регрессия, Пошаговая регрессия, CatBoost, ExtraTrees) | 11.663 |
| Классификация | Блендинг, ml | PSO (kNN, SVM, Логистич. Lasso-регрессия, XGBoost, LightGBM и др.) | 11.625 |
| Регрессия | Блендинг, chain | PSO | 11.613 |
| Классификация | Блендинг, lp | PSO / Поиск по сетке | 11.525 |
| Классификация | Блендинг, mc | Генетический алгоритм / PSO | 11.263 |
| Регрессия | Multioutput | Lasso-регрессия | 11.175 |
| Регрессия | Chain | Ridge-регрессия | 11.062 |
| Классификация | Multilabel | kNN | 10.838 |
| Ранжирование | Списочное ранжирование | MLP с ListNet@3 | 10.788 |
| Классификация | Multiclass | Логистическая регрессия (Lasso) | 10.663 |

*Обозначения:*
*mo — Multioutput, ml — Multilabel, lp — Label Powerset, mc — Multiclass,*
*MLP — многослойный перцептрон, PSO — метод роя частиц,*
*SVM — метод опорных векторов, kNN — метод k-ближайших соседей*

Таблица 13. Восстановление значений незаполненных психометрических тестов

| **Модель-регрессор** | **MICE** | **Matrix Soft Impute** | **Маски** | **Ансамбли** |
|---|---|---|---|---|
| Регрессия Lasso (L1) | 9.191 | 10.090 | 9.998 | 9.866 |
| Пошаговая регрессия | 9.608 | 9.754 | 9.978 | 10.082 |
| Линейная регр. (OLS) | 9.407 | 9.612 | 9.876 | 10.012 |
| Регрессия Ridge (L2) | 9.442 | 9.733 | 9.868 | 9.933 |
| ExtraTrees | 9.101 | 9.627 | 9.870 | 9.808 |
| Метод опорных векторов (SVR) | 9.221 | 9.622 | 9.864 | 9.760 |
| CatBoost | 9.131 | 9.766 | 9.835 | 9.461 |
| kNN | 9.372 | 9.486 | 9.830 | 9.377 |
| Random Forest | 9.518 | 9.710 | 9.819 | 9.712 |
| LightGBM | 9.372 | 9.678 | 9.686 | 9.594 |
| XGBoost | 8.769 | 9.571 | 9.267 | 9.614 |
| Constant baseline | 9.000 | 9.000 | 9.000 | 9.000 |

# A. Описание психометрических тестов

Таблица 14. Психометрические тесты: описательная статистика

| Опросник | Признак | Код | N | Mean (SD) | Median (IQR) | Min | Max |
|---|---|---|---|---|---|---|---|
| 16 факторный опросник Кеттелла | Открытость – Замкнутость | CT_1 | 993 | 9,91 (3,67) | 10 (7–12) | 0 | 19 |
| | Эмоциональная стабильность – Эмоциональная неустойчивость | CT_2 | 993 | 12,99 (4,73) | 13 (10–16) | 0 | 26 |
| | Независимость – Податливость | CT_3 | 993 | 12,84 (4,06) | 13 (10–16) | 1 | 25 |
| | Беспечность – Озабоченность | CT_4 | 993 | 12,25 (4,35) | 12 (9–15) | 2 | 25 |
| | Сознательность – Беспринципность | CT_5 | 993 | 10,61 (3,56) | 11 (8–13) | 1 | 20 |
| | Смелость – Застенчивость | CT_6 | 993 | 10,88 (5,93) | 11 (6–15) | 0 | 26 |
| | Чувственность – Твердость | CT_7 | 993 | 11,99 (3,71) | 12 (10–15) | 1 | 20 |
| | Подозрительность – Доверчивость | CT_8 | 993 | 10,72 (3,55) | 11 (8–13) | 0 | 20 |
| | Мечтательность – Практичность | CT_9 | 993 | 10,84 (3,02) | 11 (9–13) | 2 | 20 |
| | Утонченность – Простота | CT_10 | 993 | 10,06 (2,99) | 10 (8–12) | 2 | 20 |
| | Склонность к чувству вины – Спокойная самоуверенность | CT_11 | 993 | 13,99 (4,99) | 14 (10–18) | 0 | 26 |
| | Радикализм – Консерватизм | CT_12 | 993 | 10,32 (2,97) | 10 (8–12) | 0 | 20 |
| | Самостоятельность – Зависимость от группы | CT_13 | 993 | 12,60 (3,51) | 13 (10–15) | 1 | 20 |
| | Самоконтроль, сильная воля – Недостаток самоконтроля, индифферентность | CT_14 | 993 | 11,42 (3,43) | 12 (9–14) | 1 | 20 |

Продолжение на следующей странице

Таблица 14. Психометрические тесты: описательная статистика (продолжение)

| Опросник | Признак | Код | N | Mean (SD) | Median (IQR) | Min | Max |
|---|---|---|---|---|---|---|---|
| | Внутренняя напряженность – Внутренняя расслабленность | CT_15 | 993 | 14,46 (5,25) | 15 (11–18) | 0 | 26 |
| | Развитое мышление – Ограниченное мышление | CT_16 | 993 | 7,69 (2,32) | 8 (6–9) | 1 | 13 |
| Личностный опросник Айзенка | Экстраверсия | EY_1 | 1200 | 11,2 (5,38) | 11 (7–15) | 0 | 24 |
| | Психотизм | EY_2 | 1200 | 6,29 (3,31) | 6 (4–8) | 0 | 22 |
| | Нейротизм | EY_3 | 1200 | 16,25 (5,74) | 17 (12–21) | 1 | 25 |
| | Искренность | EY_4 | 1200 | 10,79 (4,36) | 11 (8–14) | 0 | 25 |
| Опросник Леонгарда-Шмишека | Гипертимность | LN_1 | 998 | 12,84 (6,47) | 12 (9–18) | 0 | 24 |
| | Дистимность | LN_2 | 998 | 14,95 (4,07) | 16 (12–18) | 2 | 24 |
| | Циклотимность | LN_3 | 998 | 14,71 (5,19) | 15 (12–18) | 0 | 24 |
| | Неуравновешенность | LN_4 | 998 | 12,27 (4,83) | 12 (8–16) | 0 | 24 |
| | Застревание | LN_5 | 998 | 11,77 (6,03) | 12 (6–15) | 0 | 24 |
| | Эмотивность | LN_6 | 998 | 14,93 (5,85) | 15 (9–18) | 0 | 24 |
| | Экзальтированность | LN_7 | 998 | 13,92 (4,6) | 14 (10–18) | 2 | 24 |
| | Тревожность | LN_8 | 998 | 13,54 (5,53) | 15 (9–18) | 0 | 24 |
| | Педантичность | LN_9 | 998 | 13,10 (4,72) | 12 (9–15) | 0 | 24 |
| | Демонстративность | LN_10 | 998 | 15,26 (5,96) | 18 (12–18) | 0 | 24 |
| Пятифакторный опросник личности | Экстраверсия – интроверсия | BF_1 | 891 | 43,54 (11,51) | 43 (35–51) | 16 | 75 |
| | Привязанность – обособленность | BF_2 | 891 | 49,76 (11,78) | 50 (42–58) | 15 | 75 |
| | Самоконтроль – импульсивность | BF_3 | 891 | 51,38 (11,34) | 51 (43–60) | 15 | 75 |

Продолжение на следующей странице

Таблица 14. Психометрические тесты: описательная статистика (продолжение)

| Опросник | Признак | Код | N | Mean (SD) | Median (IQR) | Min | Max |
|---|---|---|---|---|---|---|---|
| | Эмоциональная устойчивость – эмоциональная неустойчивость | BF_4 | 891 | 52,58 (13,96) | 54 (44–63,5) | 15 | 75 |
| | Экспрессивность – практичность | BF_5 | 891 | 54,63 (8,68) | 55 (49–61) | 15 | 75 |
| Ценностный опросник Шварца | Универсализм – НИ | SC_1 | 747 | 38,66 (10) | 40 (33–46) | -8 | 56 |
| | Безопасность – НИ | SC_2 | 747 | 25,12 (6,24) | 26 (22–29) | -5 | 35 |
| | Власть – НИ | SC_3 | 747 | 15,69 (6,58) | 16 (11–20) | -2 | 28 |
| | Гедонизм – НИ | SC_4 | 747 | 14,61 (4,8) | 15 (12–18) | -2 | 21 |
| | Самостоятельность – НИ | SC_5 | 747 | 26,65 (5,35) | 27 (24–30) | 0 | 35 |
| | Стимуляция – НИ | SC_6 | 747 | 11,36 (5,23) | 12 (8–15) | -3 | 21 |
| | Конформность – НИ | SC_7 | 747 | 17,04 (5,98) | 18 (14–21) | -4 | 28 |
| | Традиция – НИ | SC_8 | 747 | 17,36 (8,16) | 18 (12–23) | -5 | 35 |
| | Доброта – НИ | SC_9 | 747 | 24,06 (6,98) | 25 (20–29) | -3 | 35 |
| | Достижение – НИ | SC_10 | 747 | 19,49 (5,53) | 20 (16–24) | -3 | 28 |
| | Самостоятельность – ИП | SC_11 | 747 | 11,15 (3,44) | 12 (9–14) | -1 | 16 |
| | Власть – ИП | SC_12 | 747 | 4,63 (3,8) | 4 (2–8) | -3 | 12 |
| | Универсализм – ИП | SC_13 | 747 | 13,87 (5,8) | 14 (10–19) | -6 | 24 |
| | Достижение – ИП | SC_14 | 747 | 8,61 (4,54) | 9 (5–12) | -4 | 16 |
| | Безопасность – ИП | SC_15 | 747 | 10,69 (5,03) | 11 (7–14) | -5 | 20 |
| | Стимуляция – ИП | SC_16 | 747 | 5,31 (3,49) | 5 (3–8) | -3 | 12 |
| | Конформность – ИП | SC_17 | 747 | 6,43 (4,46) | 7 (3–10) | -4 | 16 |
| | Традиция – ИП | SC_18 | 747 | 4,46 (4,44) | 4 (1–7) | -4 | 16 |
| | Гедонизм – ИП | SC_19 | 747 | 7,38 (3,35) | 8 (5–10) | -3 | 12 |

Продолжение на следующей странице

Таблица 14. Психометрические тесты: описательная статистика (продолжение)

| Опросник | Признак | Код | N | Mean (SD) | Median (IQR) | Min | Max |
|---|---|---|---|---|---|---|---|
| | Доброта – ИП | SC_20 | 747 | 8,27 (4,37) | 9 (5–11) | -4 | 16 |
| Тест Голланда | Реалистический (R) | HL_1 | 1278 | 6,42 (2,33) | 6 (5–8) | 0 | 14 |
| | Исследовательский (I) | HL_2 | 1278 | 7,19 (2,18) | 7 (6–9) | 2 | 13 |
| | Артистический (A) | HL_3 | 1278 | 7,09 (2,07) | 7 (6–9) | 1 | 13 |
| | Социальный (S) | HL_4 | 1278 | 6,67 (2,04) | 7 (5–8) | 1 | 13 |
| | Предприимчивый (E) | HL_5 | 1278 | 7,04 (2,34) | 7 (5–9) | 1 | 13 |
| | Традиционный (C) | HL_6 | 1278 | 7,59 (2,78) | 8 (6–10) | 0 | 14 |