

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

ГЛУШКОВ Егор Александрович

гр. 23.M04-мм, СПбГУ

Научный руководитель: Абрамов М. В.,

к. т. н., доцент кафедры информатики СПбГУ

Консультант: Столярова В. Ф.,

к. т. н., старший преподаватель кафедры информатики СПбГУ

17 мая 2025

- Значение корректного выбора профессионального пути
- Ресурсоёмкость традиционного глубинного интервью, дистанционные методы профориентации
- Модель RIASEC (код Голланда) и её ограничения:
 - шесть типов социально-профессиональной направленности личности
 - различные вариации теста
 - культурные и социо-экономические различия респондентов
 - динамично меняющаяся ситуация на рынке профессий
- Валидация результатов тестирования, связь с другими тестами, неполнота информации — восстановление результатов теста Голланда по альтернативным тестам

Постановка задачи

Цель — разработать инструмент для автоматизации профориентации на основе определения кода Голланда по результатам психометрических тестов личности с использованием методов машинного обучения

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

Постановка задачи [2]

- *Новизна результатов исследования:* использование комбинации различных психометрических тестов для предсказания кода Голланда
- *Теоретическая значимость:* создание новых моделей машинного обучения для определения взаимосвязи психометрических тестов и кода Голланда
- *Практическая значимость:* реализация прототипа программного модуля автоматизации оценки профессиональной направленности по психологическому профилю личности

Обзор. Психометрические тесты личности

- Психологические тесты личности (количество факторов):
 - ❶ Модель Голланда RIASEC (6)
 - ❷ Опросник Леонгарда-Шмишека (10)
 - ❸ Личностный опросник Айзенка (4)
 - ❹ 16-факторный опросник Кеттелла (16)
 - ❺ Пятифакторный опросник личности («Большая пятерка»; 5)
 - ❻ Ценностный опросник Шварца (20)
- Цель тестирования — отразить некоторые черты личности человека в удобном целочисленном формате
- Модель Голланда RIASEC – 6 типов личностей, с которыми соотнесены наборы профессий:
 - реалистический (Realistic, R)
 - исследовательский (Investigative, I)
 - артистический (Artistic, A)
 - социальный (Social, S)
 - предприимчивый (Enterprising, E)
 - традиционный (Conventional, C)

Пример входных данных

Таблица 1. Пример данных психометрических тестов

	Большая Пятёрка			...	Леонгард		Голланд					
id	BF_1	BF_2	BF_3	...	LN_9	LN_10	HL_1	HL_2	HL_3	HL_4	HL_5	HL_6
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

- VK Mini Apps «Психологические тесты»
- Анонимизированные данные 1278 пользователей: 339 – полные, 939 – заполнены данные по 4 или 5 тестам
- Обработка данных: валидация, заполнение пропусков, стандартизация, уменьшение размерности PCA

Методы решения

Методы решения задачи по определению кода Голланда:

- Многоцелевая регрессия
- Классификация
- Ранжирование

Реализация вычислительного эксперимента:

- R (data.table, R6Class), Python (PyTorch)
- Стенд: R Shiny

Регрессия. Подходы

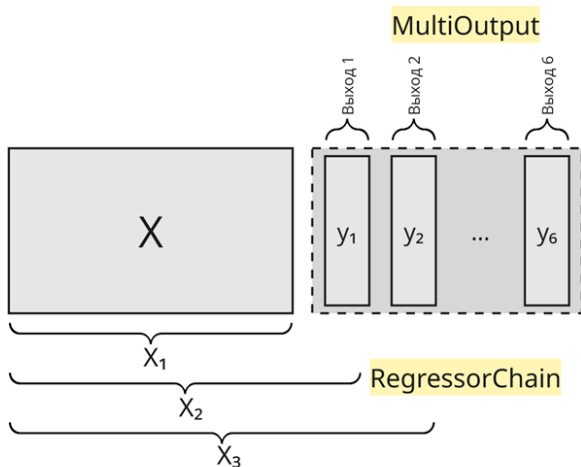


Рисунок 1. Подходы к решению задачи многоцелевой регрессии

Регрессия. Метрики качества

- Усреднённая среднеквадратичная ошибка (RMSE)
- Усреднённый C-индекс: мера согласованности (чем больше, тем более схожи профили личности)
- 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:

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

- Пример: $C([6, 10, 4, 5, 8, 9], [6, 6, 7, 9, 6, 8]) =$

$$C(\text{ICE}, \text{SCA}) = 3 \cdot 1 + 2 \cdot 3 + 1 \cdot 1 = 10$$

Регрессия. Результаты (С-индекс)

Таблица 2. Сравнение моделей по метрике С-индекс

Модель	МО	МО 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, регул-я)	10.462			
MLP	10.275			

* МО — Multioutput, PCA — метод главных компонент (уменьшение размерности)

Регрессия. Результаты (RMSE)

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			

* MO — Multioutput, PCA — метод главных компонент (уменьшение размерности)

Анализ важности признаков

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

Регрессия. Ансамблевые модели

- **Ансамблевые модели:**

- Стекинг
- Линейный блендинг (взвешенное ансамблирование)

- Методы подбора весов для блендинга:

- Равные веса всех моделей
- Вектор Шэпли
- Частичный перебор по сетке
- Квадратичная оптимизация (QP)
- Генетический алгоритм (GA)
- Метод роя частиц (PSO)
- Координатный спуск

- Опционально использовалось уменьшение размерности (PCA)

Регрессия. Ансамблирование регрессоров – результаты

Таблица 5. Сравнение методов подбора весов ансамбля регрессионных моделей

Метод подбора весов	МО	МО 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 регрессия		10.887	
Lasso, Ridge, LightGBM, CatBoost, RF	Линейная регрессия		10.688	

* МО — 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	Extra Trees	
PSO	0.432	0.327	0.150	0.091	11.663

Распределение С-индекса предсказаний

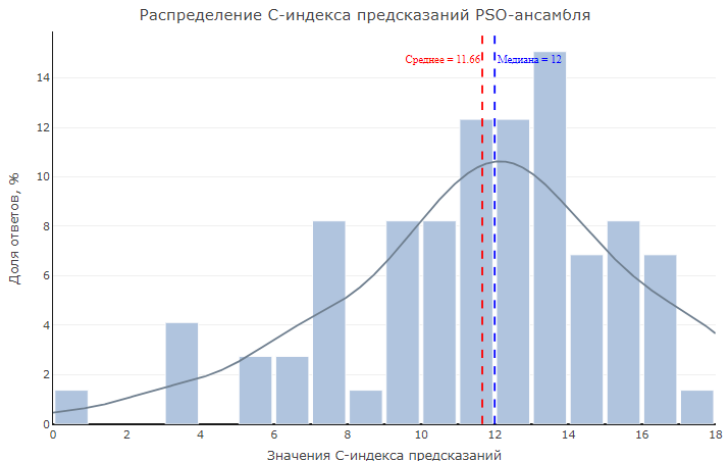


Рисунок 2. Распределение С-индекса предсказаний PSO-модели

Классификация. Подходы

Таблица 7. Подходы к классификации

Тип классификации	Модель	Значение	Пример
Многоклассовая (multiclass)	1 классификатор на 6 классов	3 кода из 1 буквы	R, S, E
Многометочная (multilabel)	6 бинарных классификаторов	Булевый вектор из 6 элементов с 3 True	[T, F, F, T, T, F]
Label powerset	1 классификатор на 20 классов (порядок не важен)	3-буквенный код	RSE

- Верхняя триада — три наиболее значимых кода
- Метрики: Top-K accuracy, C-индекс
- Опционально уменьшение размерности (PCA)

Классификация. Сравнение подходов

Таблица 8. Сравнение подходов к классификации (Top-K accuracy)

Модель	Multiclass			Multilabel			Label Powerset		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
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

Классификация. Сравнение моделей

Таблица 9. Сравнение классификаторов

Классификатор	Подход	С-индекс	Тор-1	Тор-2	Тор-3
kNN	Multilabel	10.838	1.000	0.763	0.113
Логистическая регр. (L1)	Multiclass	10.663	1.000	0.700	0.163
XGBoost	Multiclass	10.638	1.000	0.700	0.113
Логистическая регр. (L2)	Multiclass	10.500	1.000	0.700	0.150
Наивный Байес	Multilabel	10.350	0.988	0.700	0.150
Extra Trees	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

Ансамблирование классификаторов – результаты

Таблица 10. 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 11. Весовые коэффициенты моделей и C-индекс

Метод подбора весов	Веса моделей									C-инд
	kNN	SVM	Logit	L1	XGBoost	LightGBM	NaiveBayes	ExtraTrees	RF	
PSO	0.291	0.164	0.191		0.183	0.151	0.009	0.007	0.004	11.625

Ранжирование

- Списочное ранжирование — Listwise Learn-to-Rank
- Задание скоринговой функции (MLP, Deep and Cross Network, Transformer) и функции потерь для списков (NDCG, ApproxNDCG, LambdaRank, ListNet@k)

Таблица 12. Сравнение моделей ранжирования

Функция потерь	С-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

Восстановление пропусков в случае неполноты данных

Таблица 13. Восстановление значений незаполненных психометрических тестов

Модель-регрессор	MICE	Matrix Soft Impute	Маски	Линейный блендинг
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

Сводные результаты по типам задач

Таблица 14. 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	Multiooutput	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
Регрессия	Восстановление пропусков	Matrix Soft Imputation + Lasso-регрессия	10.090

Обозначения:

mo — Multiooutput, ml — Multilabel, lp — Label Powerset, mc — Multiclass,

MLP — многослойный перцептрон, PSO — метод роя частиц,

SVM — метод опорных векторов, kNN — метод k-ближайших соседей

Итоговая последовательность этапов

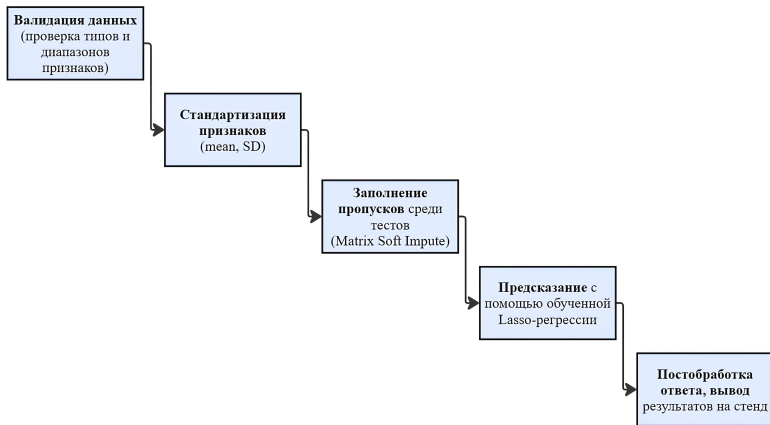


Рисунок 3. Итоговая последовательность вычислительных этапов для определения кода Голланда по психометрическим тестам при неполных данных

Демонстрация прототипа инструмента

Предсказание кода Голланда по результатам психометрических тестов

☐ Тест 16-факторный опросник Кеттелла (16 факторов)

☒ Тест Личностный опросник Айзенка (4 фактора)

1. Экстраверсия

0

Допустимо: от 0 до 25

2. Психотизм

25

Допустимо: от 0 до 25

3. Нейротизм

25

Допустимо: от 0 до 25

4. Искренность

11

Допустимо: от 0 до 25

☐ Тест Опросник Леонгарда-Шмишека (10 факторов)

☒ Тест Пятифакторный опросник личности (5 факторов)

☐ Тест Ценностный опросник Шварца (20 факторов)

Подсчитать

Результаты прогноза

Прогноз сделан на основе результатов следующих тестов:

- Личностный опросник Айзенка
- Пятифакторный опросник личности

Коды Голланда:

- Наиболее вероятные: I (55.3%), C (19.4%), R (13.4%)
- Менее вероятные: S (8.5%), A (2.4%), E (1.5%)

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).
Любит анализировать данные, исследовать гипотезы и решать интеллектуальные задачи. Стремится к научным открытиям и пониманию сложных систем. Примеры: учёный , программист, биолог, земек.
- 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 профессию, связанные с физическим трудом или природой. Примеры: инженер , механик, строитель, фермер.

Рисунок 4. Демонстрация стенда¹.

¹GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности. URL: https://github.com/ExP98/Diploma_Holland (дата обращения: 17.05.2025).

Демонстрация прототипа инструмента [2]

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. Эзепрессивность - 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 Голланде:

- Наиболее вероятные: I (50,4%), С (18%), Я (16,3%)
- Менее вероятные: S (10,5%), А (2,8%), Е (2%)

Обозначения:

X (Y%), где X – код Голланды, соответствующий типу личности,

Y – степень уверенности, что данный код Голланды входит в верхнюю триаду.

Ваши типы личности:

- **I (Исследовательский).**
Любит анализировать данные, исследовать гипотезы и решать интеллектуальные задачи. Стремится к научным открытиям и пониманию сложных систем. Примеры: учёный, программист, биолог, химик.
- **II (Конвенциональный).**
Предпочитает четкие инструкции, структуру и работу с цифрами/документами. Ценит аккуратность и системный подход. Примеры: бухгалтер, администратор, налоговый инспектор, логист.
- **III (Реалистичный).**
Предпочитает практические задачи, работу руками и с техникой. Часто выбирает профессии, связанные с физическим трудом или техникой. Примеры: инженер, механик, строитель, фермер.

Подсчет выполнен

Рисунок 5. Демонстрация стенда² [2]

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

Результаты

- ❶ Проанализированы существующие подходы к определению кода Голланда, поставлены задачи:
 - регрессии (multioutput, chain)
 - классификации (multiclass, multilabel, label powerset)
 - линейного блендинга базовых моделей
 - ранжирования
- ❷ Реализованы модели для предсказания кодов Голланда: лучшие результаты у моделей линейного блендинга на основе метода роя частиц (С-индекс):
 - Ансамбль multioutput-регрессоров: Lasso-регрессия, пошаговая регрессия, CatBoost, ExtraTrees (11.663)
 - Ансамбль multilabel-классификаторов: kNN, SVM, логистическая Lasso-регрессия, XGBoost, LightGBM и др. (11.625)
 - Показано преимущество «классических» методов перед нейросетевыми
- ❸ Реализованы алгоритмы восстановления результатов психометрических тестов: лучший – метод мягкой импутации в сочетании с Lasso-регрессией (10.09)
- ❹ Создан прототип инструмента для определения профориентационных предпочтений: стенд на основе R Shiny