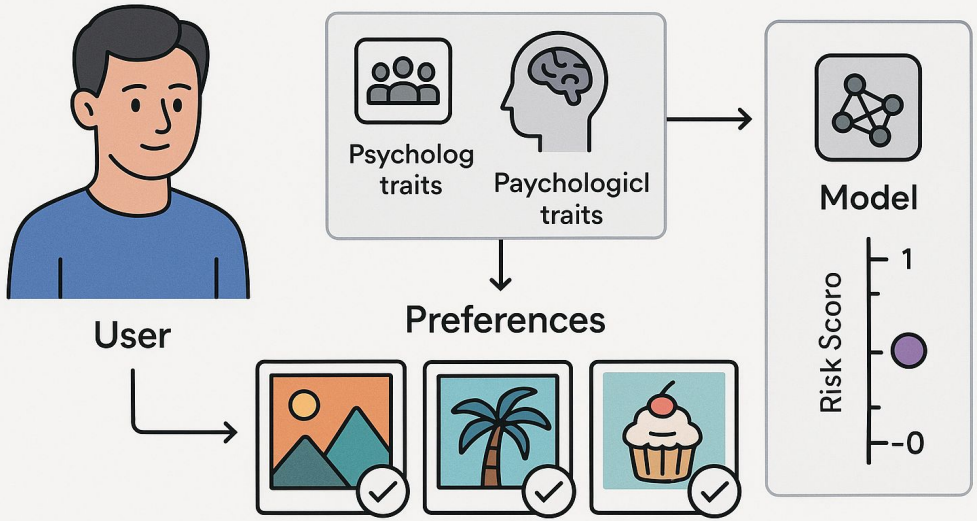
Risk Profiling using LLM

Alina Salimova, Higher School of Economics

Предлагаемый подход



Индивидуальный портрет и риск-профиль формируются на основе двух уровней признаков:

- **Демографические характеристики** (возраст, пол, образование, доход)
- **Скрытые психологические паттерны**, которые предлагается извлекать через визуальные предпочтения

Генерация изображений

Google Gemini API

Четыре ключевых оси:

- Безопасность vs Свобода
- Активность vs Расслабление
- Порядок vs Хаос
- Социальность vs Интроверсия

Визуальный стиль: мультяшный реализм во вселенной «Шрека»

16 сцен, сочетание значений по четырём осям (1-1-2-2 и тд)

Работа с реальными данными SCF 2022

- **T-pro-it-1.0, GPT-4o**
- **Задача**: представить, что LLM — реальный человек из выборки
- На основе профиля выбирать подходящие картинки
- Сложная структура данных

Категориальные переменные:

HHSEX, RACECL, EDUC, MARRIED, SPENDMOR, YESFINRISK, NOFINRISK, BFINPLAN, LATE60, BNKRUPLAST5, ANYPEN, HTRAD

Числовые переменные:

AGE, INCOME, WAGEINC, INTDIVINC, KGINC, NORMINC, CHECKING, SAVING, STOCKS, BOND, EQUITY, RETQLIQ, VEHIC, HOUSES, IRAKH, ASSET, NETWORTH, HOMEEQ, MRTHEL, HELOC, CCBAL, VEH_INST, DEBT, DEBT2INC, LEVRATIO

Неконсистентность при генерации

1ый запуск:

	row_index	investor_description	first_choice	second_choice	third_choice
0	20310	Female, age 53. White/Caucasian. education: 12...	2-2-1-1	2-1-1-2	1-2-1-2
1	22480	Female, age 59. White/Caucasian. education: as...	1-2-1-2	2-2-1-2	2-1-1-2

2ой запуск:

	row_index	investor_description	first_choice	second_choice	third_choice
0	20310	Female, age 53. White/Caucasian. education: 12...	1-2-1-2	2-2-1-2	2-2-2-2
1	22480	Female, age 59. White/Caucasian. education: as...	1-2-1-2	2-2-1-2	2-1-1-2

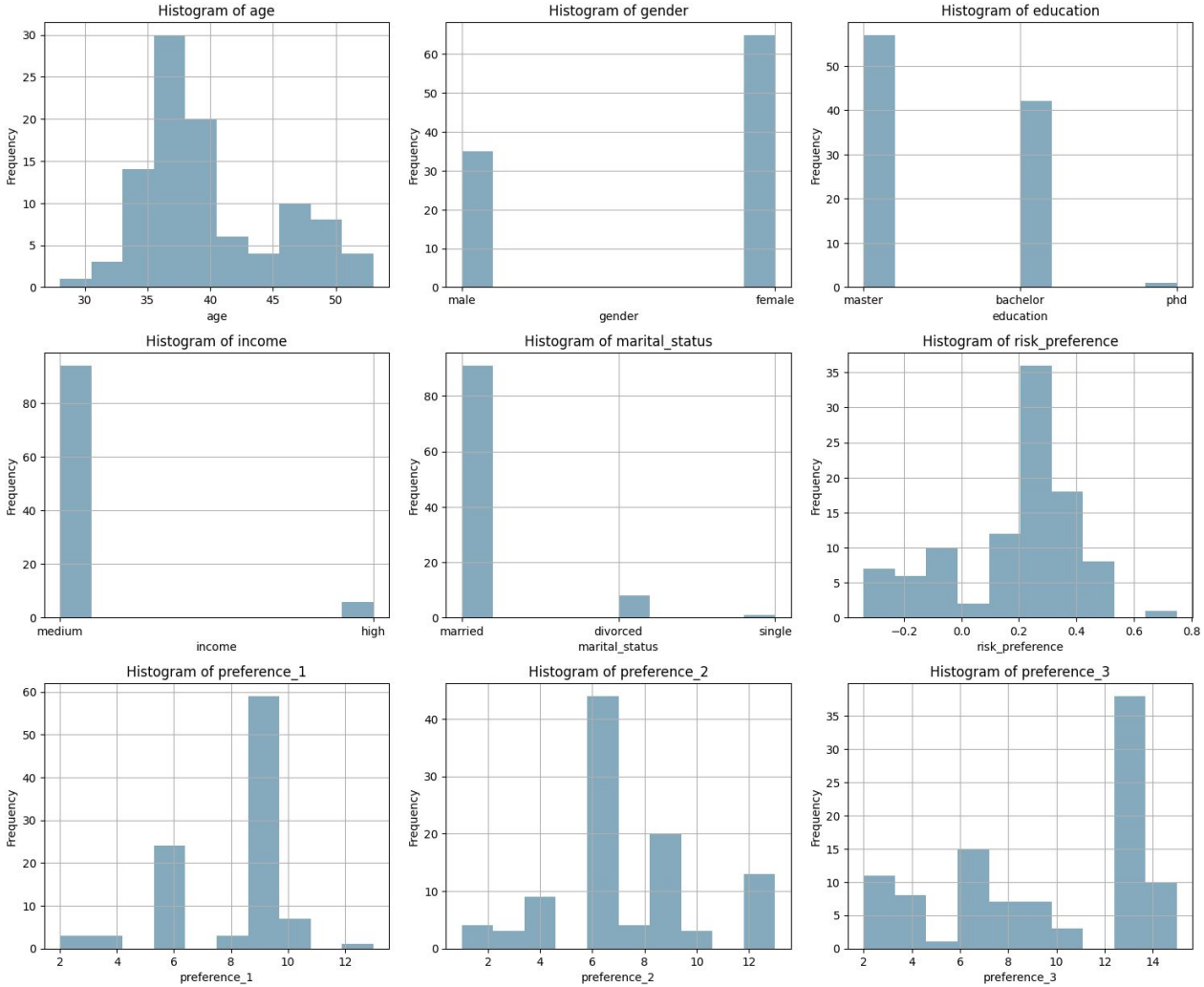
Переменная	Распределение
Возраст (18–80)	Сэмплирование по весам: 18–20: 4%, 21–45: 70%, 46–65: 9%, 66–80: 3%
Пол	Бернулли: 55% мужчин / 45% женщин + редкие «другое»
Образование	Высшее – 62%, среднее спец. – 35%, школа – 3%
Доход	Низкий – 10%, средний – 60%, высокий – 30%
Семейный статус	Зависит от возраста: - 18–29: single 60%, married 30%, divorced 5%, widowed 5% - 30–45: married 65%, single 20%, divorced 10%, widowed 5% - 46+: married 50%, divorced 20%, widowed 10%, single 20%
Риск-профиль	Из нормального $N(\mu = -0.1; \sigma = 0.5)$, обрезанного до диапазона $[-1; 1]$ и округленного до сотых

Эмпирические источники:

“Профиль инвестора в России”, 2024

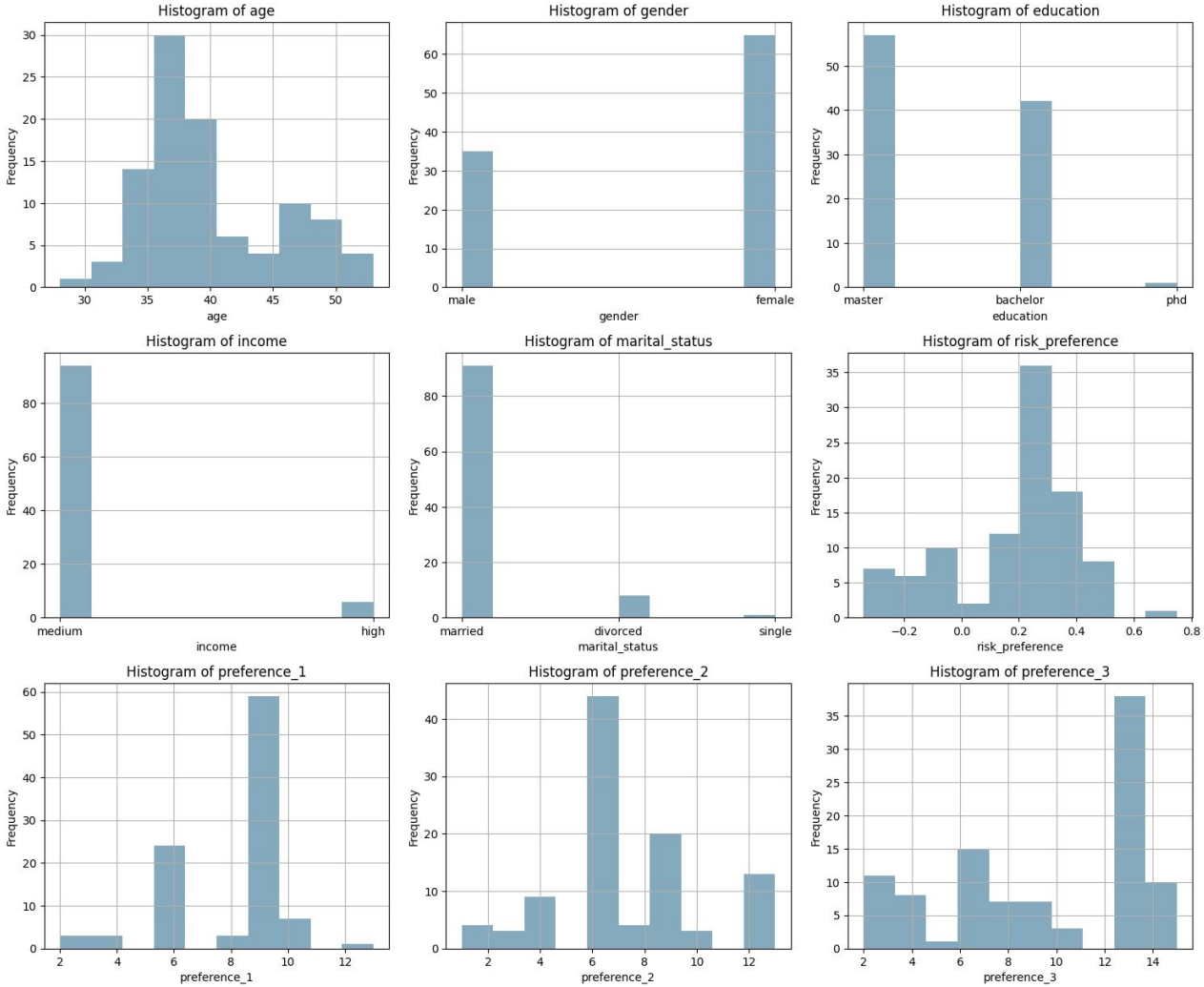
«Determinants of Private Investors’ Behavior on Russian Stock Market», 2020

«The Impact of Financial Literacy on the Choice of Financial Instruments by Private Investors in Russian Conditions», 2025



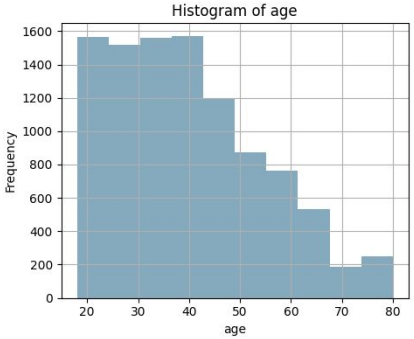
0. LLM-sampled

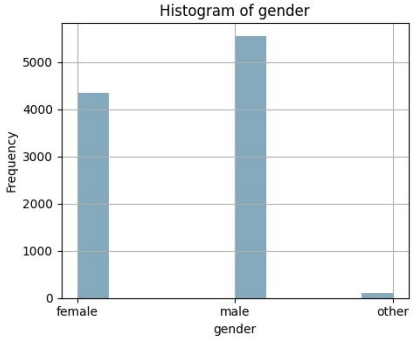
LLM придумывает личность (например, «30-летняя учительница, замужем, средний доход») и выбирает подходящие изображения

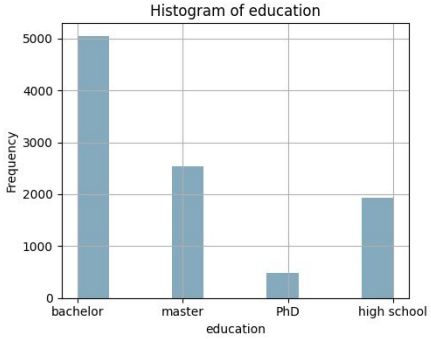


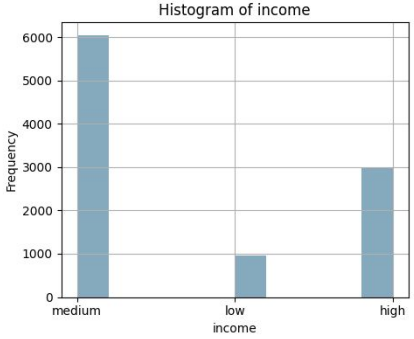
1. LLM-sampled, recommended distribution

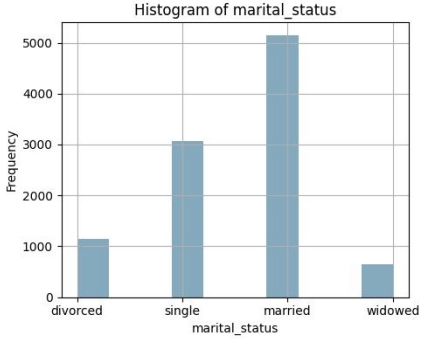
- В prompt явно описаны распределения (например, возраст 30–45 с вероятность 0.6)
- LLM представляет себя этим человеком и выбирает изображения на основе этих данных

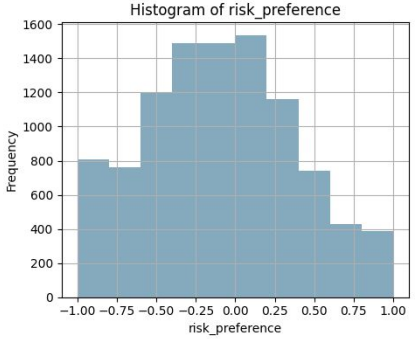


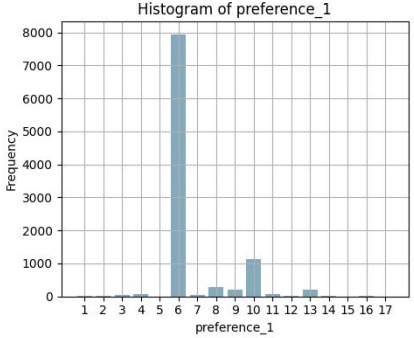


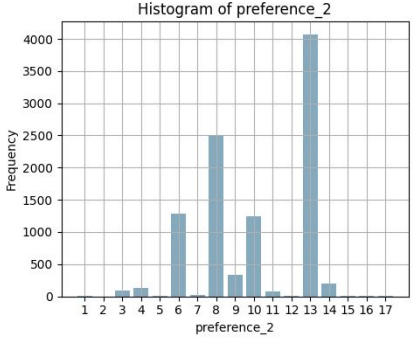


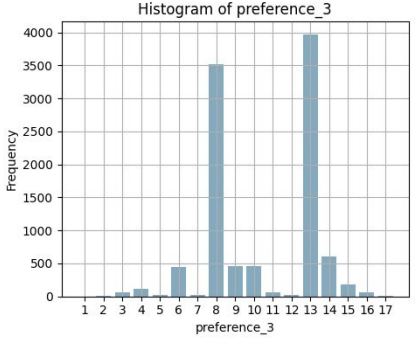












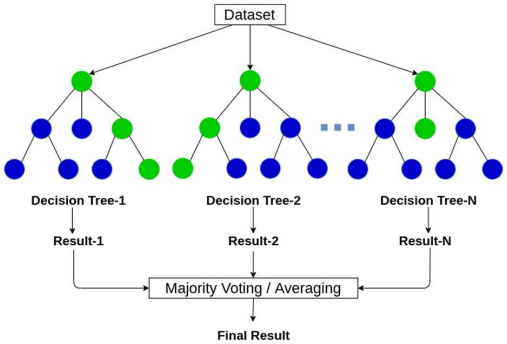
2.Pre-sampled

- Сначала генерируются профили программно, используя заданные распределения
- Затем полученный профиль подается в LLM вместе с инструкцией «представь, что это ты, выбери картинки»

Models

Baseline

0. Baseline (mean prediction)



Random Forest

1. RF (socdem only)
Только соцдемография. Почти не даёт прироста качества.

2. RF (socdem + weighted meta)
Соцдемография + агрегированные (взвешенные) мета-признаки из описаний

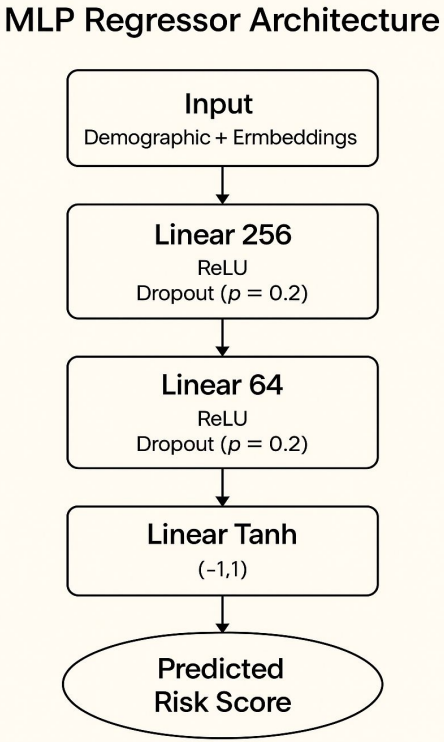
3. RF (socdem + one-hot meta)
Соцдемография + one-hot мета-признаки из аннотаций.

4. RF (meta only)
Только мета-признаки без соцдемографических данных

5. RF (text embeddings)
Эмбеддинги текстов дают заметное улучшение точности

6. RF (img embeddings)
Эмбеддинги изображений — близкий результат к текстовым

Models



MLP

7. MLP (text + socdem)
Нейросеть на текстовых эмбеддингах и соцдем-признаках

8. MLP (img + socdem)
Нейросеть на эмбеддингах изображений и соцдемографии.

9. MLP (socdem clusters)
Соцдемография сначала кластеризуется, затем обучается MLP

Results

	model_name	mse	r2	train_time	inference_time
0	Baseline (mean prediction)	0.2395	0.0000	0.0011	0.0000
1	RandomForest (socdem only)	0.2397	-0.0039	0.3062	0.0120
2	RandomForest (socdem + weighted meta)	0.2325	0.0266	0.5402	0.0130
3	RandomForest (socdem + one-hot meta)	0.2320	0.0284	0.7132	0.0161
4	RandomForest (meta only)	0.2320	0.0284	0.3075	0.0137
5	RandomForest (text embeddings)	0.2303	0.0356	30.7492	0.0177
6	RandomForest (img embeddings)	0.2303	0.0356	46.5691	0.0241
7	MLP on text+socdem	0.2279	0.0459	21.5193	0.0156
8	MLP on img+socdem	0.2293	0.0400	23.8672	0.0189
9	MLP with clustering on socdem	0.2298	0.0378	6.4228	7.3422

Inference

Investor Risk Score Estimator

Select 3 images in order of your preference:

ID 1

Select ID 1

ID 2

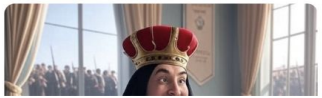
Select ID 2

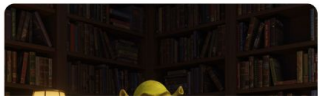
ID 3

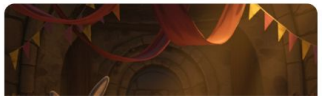
Select ID 3

ID 4

Select ID 4







Inference

ID 9

ID 10

ID 11

ID 12

- [ ] Select ID 9
- [x] Select ID 10
- [x] Select ID 11
- [ ] Select ID 12

ID 13

ID 14

ID 15

ID 16

- [ ] Select ID 13
- [x] Select ID 14
- [ ] Select ID 15
- [ ] Select ID 16

✅ 3 images selected. You can proceed!

Inference

ID 13

ID 14

ID 15

ID 16

☐ Select ID 13 ☑ Select ID 14 ☐ Select ID 15 ☐ Select ID 16

✅ 3 images selected. You can proceed!

Set preference order for selected images:

Preference 1

14

10

11

14

Inference

Sociodemographic Information:

Age

30

18 100

Gender

female

Education

PhD

Income

high

Marital Status

divorced

Calculate Risk Score

Predicted Risk Score: -0.102

Thank you!