

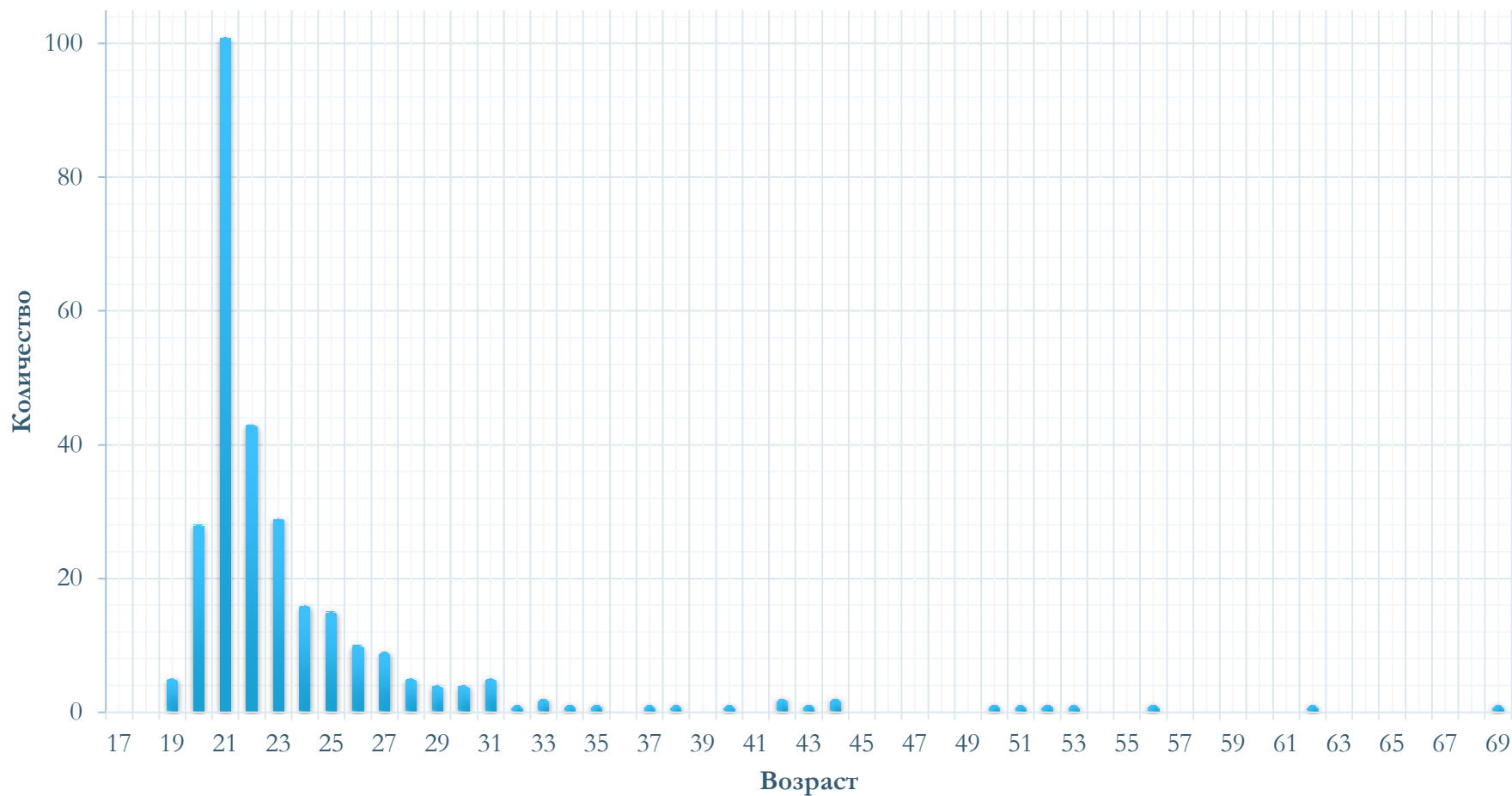
ОПРЕДЕЛЕНИЕ ВОЗРАСТА ПОЛЬЗОВАТЕЛЕЙ FACEBOOK

Кондратьев Михаил
Филипенко Максим
Тарабан Илья
Людвиченко Виталий
Димитриев Станислав


РАЗМЕР ВЫБОРКИ

294

РАСПРЕДЕЛЕНИЕ ВОЗРАСТА



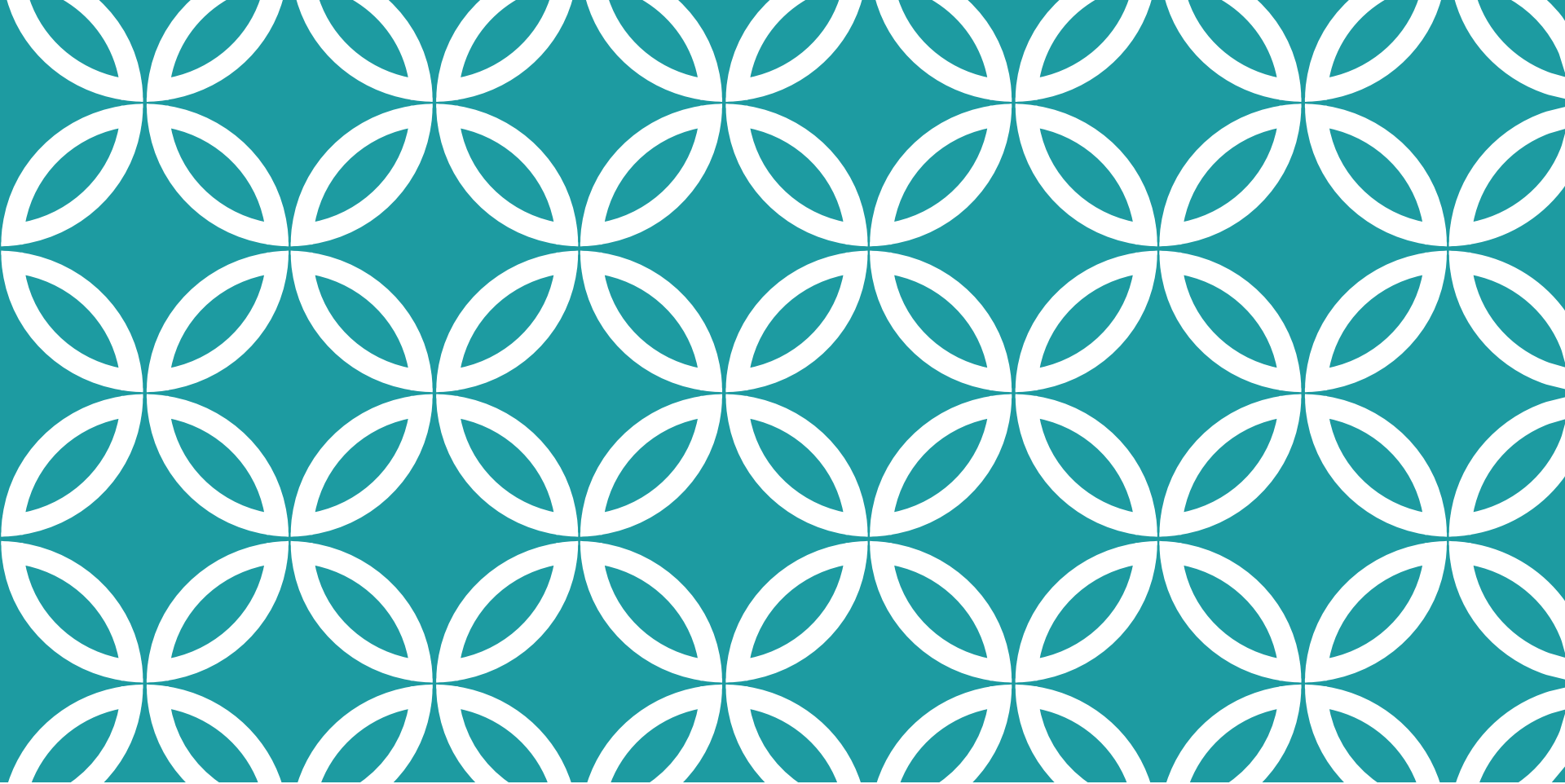
МЕТОД ТЕСТИРОВАНИЯ АЛГОРИТМОВ

$$MSE = \frac{1}{N} \sum (y(x_i) - t_i)^2, \quad \boxed{RMSE} = \sqrt{MSE}$$


$$\Rightarrow \boxed{MAE} = \frac{1}{N} \sum |y(x_i) - t_i|$$

$$\Rightarrow \boxed{RSE} = \frac{\sum (y(x_i) - t_i)^2}{\sum (t_i - \bar{t})^2}$$

$$\Rightarrow \boxed{correlation} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$



CLASSIFICATION AND REGRESSION TREE

Кондратьев Михаил

CLASSIFICATION AND REGRESSION TREE

Использованные признаки, в порядке убывания важности:

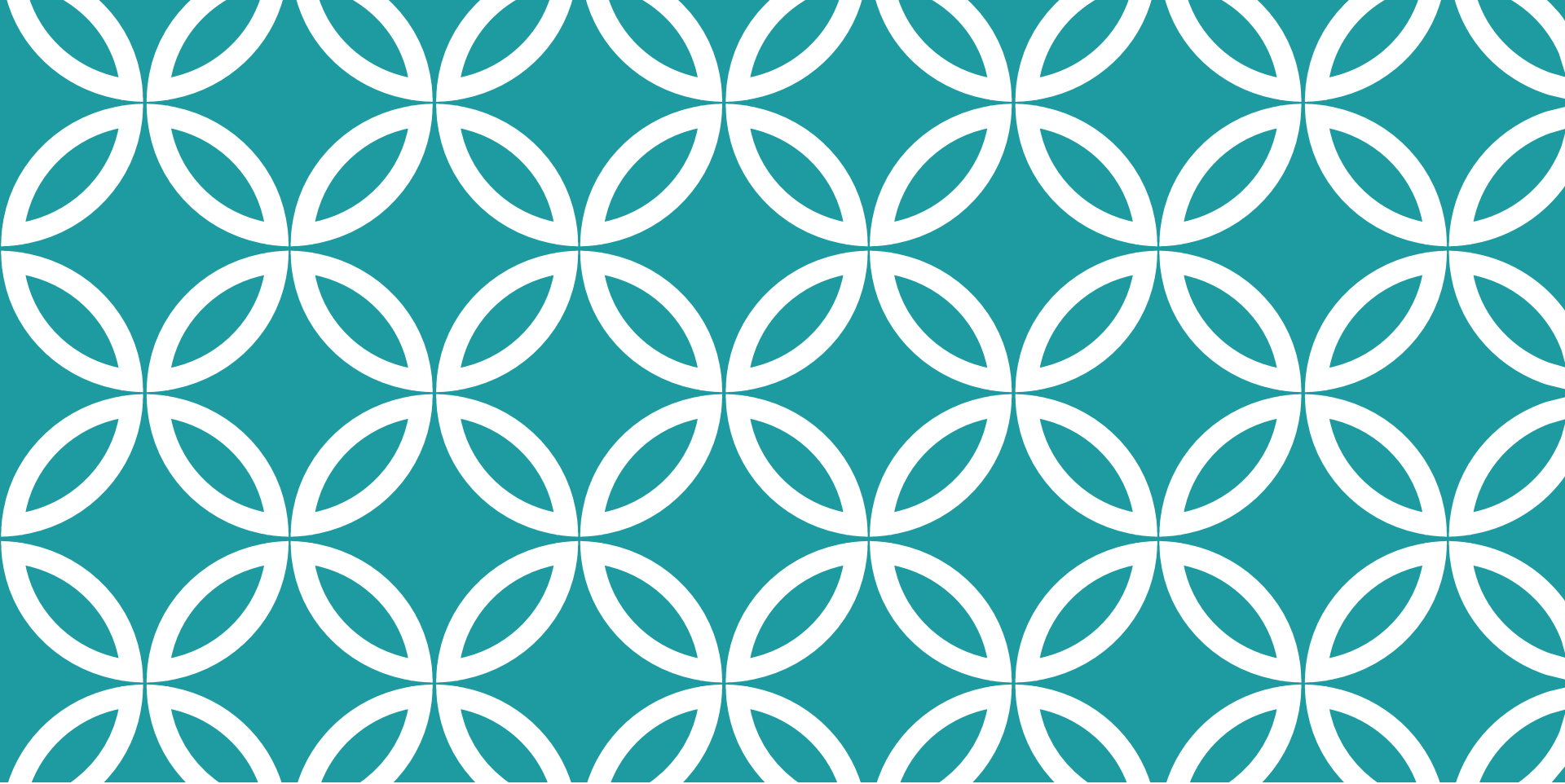
	0	1	2	3	4	5	6	7	8	9
▪ Количество мест работы	177	78	24	8	3	3	1	0	0	0
▪ Количество учебных заведений	31	41	126	48	17	15	9	5	1	1
▪ Семейное положение										
▪ Не указано: 184										
▪ Single: 56										
▪ It's complicated: 1										
▪ In a relationship: 34										
▪ In an open relationship: 0										
▪ Engaged: 4										
▪ In a domestic partnership: 1										
▪ Married: 13										
▪ Widowed: 1										
▪ <u>Совпадение/несовпадение текущего города и города в котором пользователь родился: 81/213</u>										
▪ Количество языков										
▪ Наличие цитат у пользователя										
▪ Количество видов спорта										

Переобучение

CLASSIFICATION AND REGRESSION TREE

Средние значения метрик при кросс-валидации:

- RMSE: 5.8
- MAE: 3.8
- RSE: 2.1
- Correlation: 0.32



ЛИНЕЙНАЯ РЕГРЕССИЯ

Людвиченко Виталий

ЛИНЕЙНАЯ РЕГРЕССИЯ

Детали реализации

- Используемая метрика: L_2
- Алгоритм оптимизации: наискорейший градиентный спуск

$$x_{n+1} = x_n - \alpha_n \mathbf{grad}(x_n)$$

$$\alpha_n = \operatorname{argmin}_{\alpha} f(x_n - \alpha \mathbf{grad}(x_n))$$

ЛИНЕЙНАЯ РЕГРЕССИЯ

Использованные признаки:

- Пол
- Семейное положение
- Время, прошедшее с поступления в последнее учебное заведение
- Время, прошедшее с первого трудоустройства

ЛИНЕЙНАЯ РЕГРЕССИЯ

Результаты без кросс-валидации

MAE	RMSE	RSE	Correlation
3.08	0.31	2.25	0.42

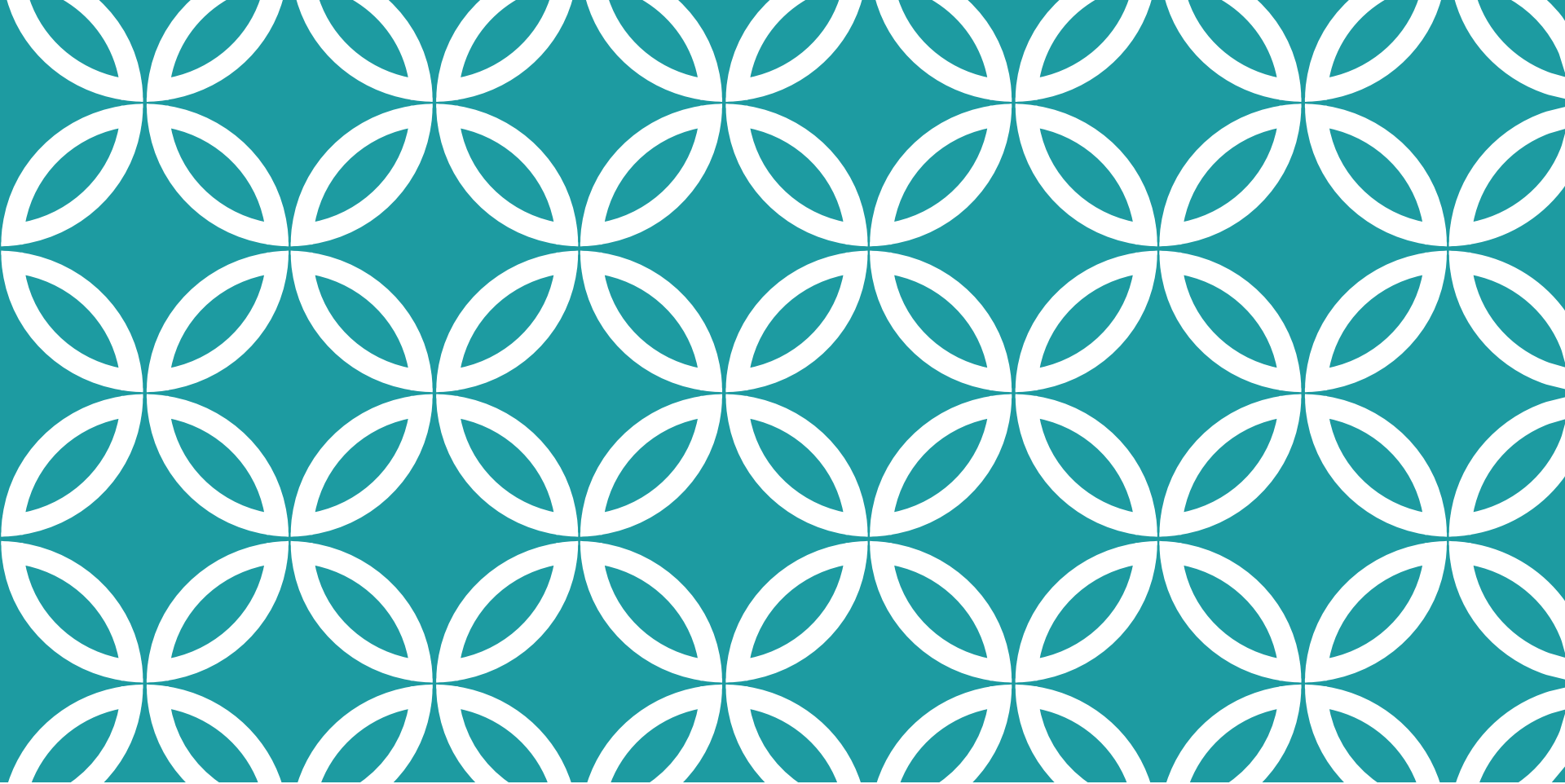
Результаты на кросс-валидации (cv = 5)

MAE	RMSE	RSE	Correlation
3.24	0.71	2.60	0.47

ЛИНЕЙНАЯ РЕГРЕССИЯ

Веса в обученной модели

Признак	Вес
one	20,28
relationship=Widowed	11,17
relationship=Married	5,85
relationship=In a domestic partnership	3,02
empty_last_education	2,97
relationship=Engaged	-1,02
last_education	0,83
relationship=It's complicated	0,73
relationship=In a relationship	0,63
empty_first_work	0,58
first_work	0,52
relationship=Single	-0,43
empty_relationship	0,31



NAIVE BAYES CLASSIFIER

Филипенко Максим

NAIVE BAYES CLASSIFIER

Использованные признаки для регрессии:

- место рождения
- место проживания
- спортивные занятия
- работа
- специальность обучения
- школа
- год окончания школы
- семейное положение

Задача регрессии для байесовских классификаторов была сведена к задаче классификации. Были выбраны интервалы 0-16, 16-18, 18-20, 20-22, 22-24, 24-26, 26-28, 28-30, 30-32, 33+

NAIVE BAYES CLASSIFIER

Результаты точности предсказания при $\alpha = [0.1, 0.5, 1.0, 1.5, 2.0, 5.0]$:

- MultinomialNB(alpha=0.1): Mean accuracy: 0.19
- BernoulliNB(alpha=0.5): Mean accuracy: 0.51
- BernoulliNB(alpha=2.0): Mean accuracy: 0.49
- GaussianNB(): Mean accuracy: 0.26

Здесь и далее была использована кросс-валидация с разбиением на пять групп ($cv = 5$)

NAIVE BAYES CLASSIFIER

Другие метрики для регрессии:

- MultinomialNB(alpha=0.1):

- Average RSE: 2.52
- Average MSE: 62.76
- Average RMSE: 5.59
- Average MAE: 6.34
- Average correlation: 0.23

- BernoulliNB(alpha=0.5):

- Average RSE: 1.31
- Average MSE: 32.54
- Average RMSE: 5.69
- Average MAE: 2.99
- Average correlation: 0.15

NAIVE BAYES CLASSIFIER

- GaussianNB():
 - Average RSE: 1.59
 - Average MSE: 39.84
 - Average RMSE: 6.27
 - Average MAE: 4.25
 - Average correlation: 0.18

NAIVE BAYES CLASSIFIER

Использованные признаки для определения пола :

- first_name
- last_name

Результаты при $\alpha = [0.1, 0.5, 1.0, 1.5, 2.0, 5.0]$:

- MultinomialNB(alpha=0.5): mean accuracy: 0.77
- BernoulliNB(alpha=0.1): mean accuracy: 0.78 (лучший)
- BernoulliNB(alpha=5.0): mean accuracy: 0.66
- GaussianNB(): mean accuracy: 0.50

NAIVE BAYES CLASSIFIER

Мало признаков — не интересно

Добавлены:

- название школы
- направление обучения
- спортивные увлечения
- название работы

Результаты:

- в целом, стало хуже
- при альфа больше 0.5 точность ухудшается
- MultinomialNB(alpha=0.1): mean accuracy: 0.73
- MultinomialNB(alpha=0.5): mean accuracy: 0.73
- MultinomialNB(alpha=1.0): mean accuracy: 0.71
- BernoulliNB(alpha=0.5): mean accuracy: 0.72
- BernoulliNB(alpha=1.0): mean accuracy: 0.68

NAIVE BAYES CLASSIFIER

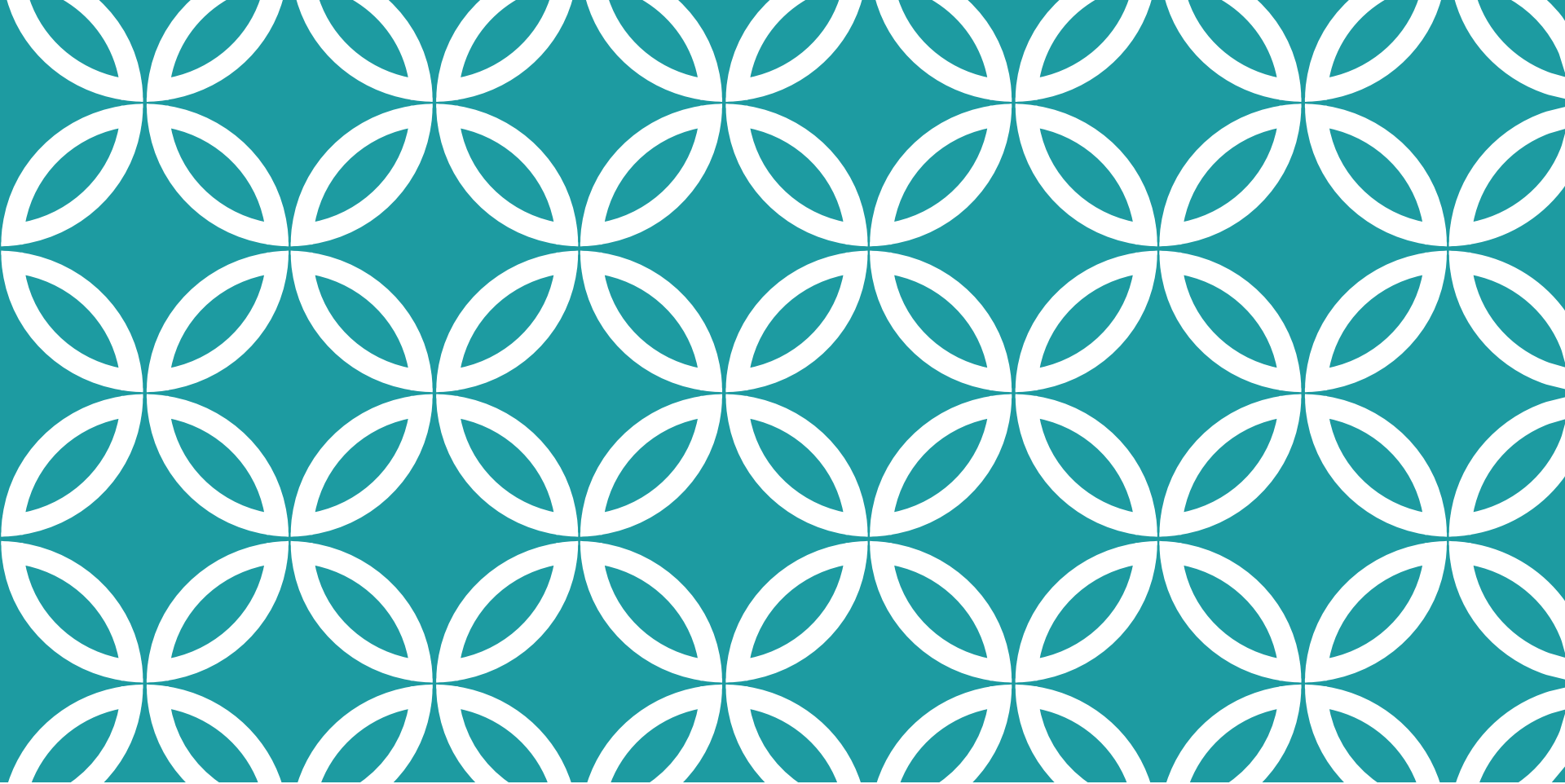
А что если удвоить присутствие `first_name` и `last_name` в выборке?

Результаты с продублированными `first_name` и `last_name`:

- MultinomialNB(alpha=0.1): mean accuracy: 0.75
- MultinomialNB(alpha=0.5): mean accuracy: 0.74
- MultinomialNB(alpha=1.0): mean accuracy: 0.74
- BernoulliNB(alpha=0.5): mean accuracy: 0.72
- BernoulliNB(alpha=1.0): mean accuracy: 0.68

Нечестность:

- слишком много мусора в данных



SUPPORT VECTOR MACHINE

Тарабан Илья

SUPPORT VECTOR MACHINE

Использованные признаки:

1. Пол
2. Статус отношений
3. Количество учебных заведений
4. Количество мест работ
5. Количество книг, фильмов, музыки
6. Количество постов
7. Частота написания постов
8. Типы использованных устройств

SUPPORT VECTOR MACHINE

Получившиеся оценки:

$MAE = 3.03$ при Линейной модели

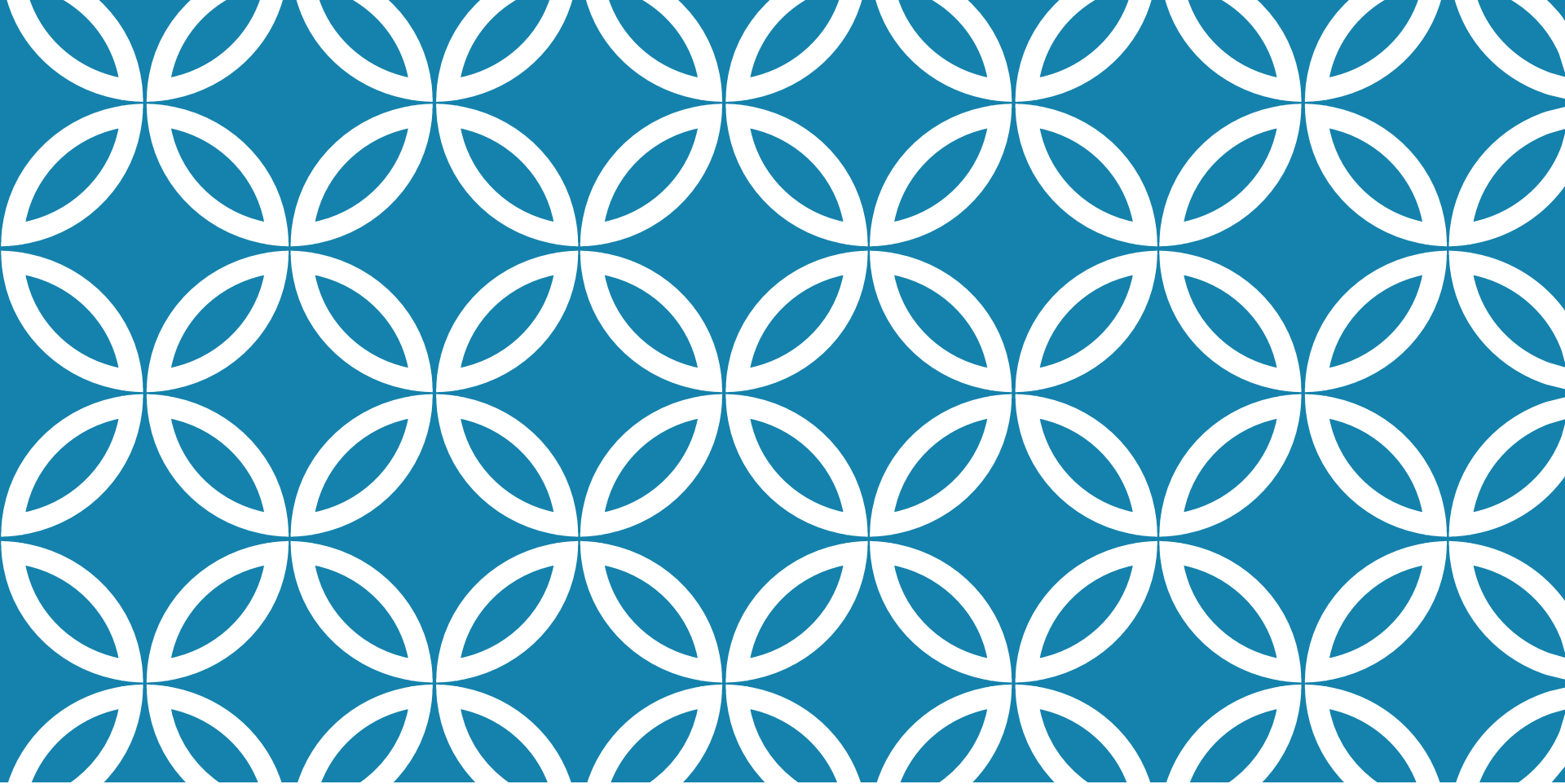
$RMSE = 5.50$ при RBF модели

$RSE = 1.27$ при RBF модели

$Correlation = 0.31$ при Линейное модели

ИТОГОВОЕ СРАВНЕНИЕ

	CART	Linear	NBC(Bernoulli)	SVM
MAE	3.8	3.24	2.99	3.03
RMSE	5.8	0.71	5.69	5.50
RSE	2.1	2.60	1.31	1.27
Correlation	0.32	0.47	0,23	0.31



СПАСИБО ЗА
ВНИМАНИЕ