Data learning

Курс “Машинное обучение”

Лабораторная работа

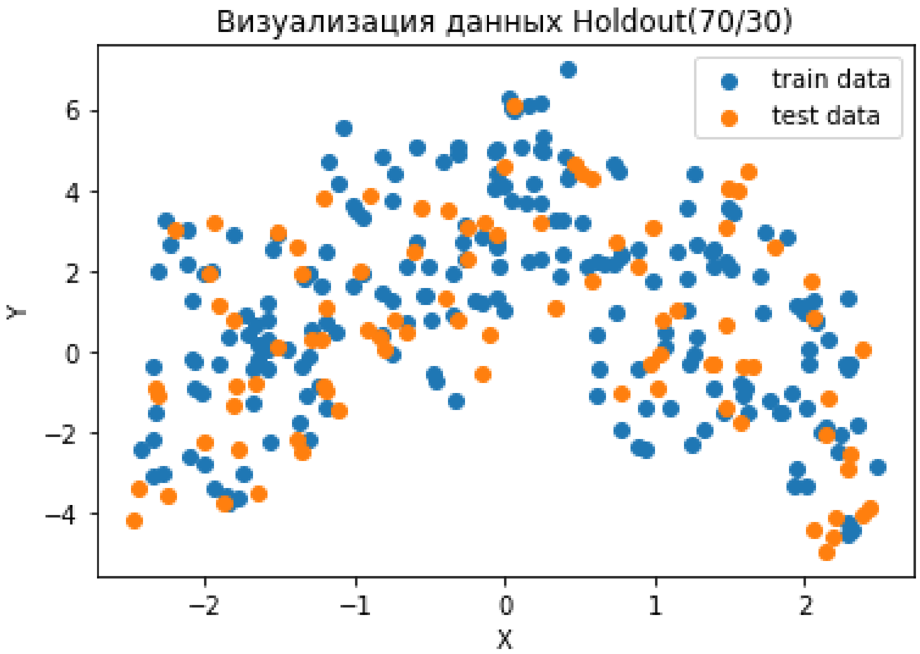
Bias-Variance decomposition

Глушков А.Е., М21-524

Вариант 1-01

2021

Исходные данные



	x	y
count	300.000000	300.000000
mean	-0.000178	0.907282
std	1.437335	2.540319
min	-2.476800	-4.940000
25%	-1.293675	-0.912722
50%	-0.048087	0.912895
75%	1.276550	2.719525
max	2.480700	6.987700

```
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   x       300 non-null    float64
 1   y       300 non-null    float64
```

Используемые методы и формулы

Simple Linear Regression: $Y|_x = \beta_0 + \beta_1 x + \varepsilon(x)$

Regression function: $\varphi(x) = \mathrm{M}[Y|_x] = \beta_0 + \beta_1 x$

Regression models class: $\mathrm{h}(x) = \beta_0 + \sum_{i=1}^{m} \beta_i x^i$

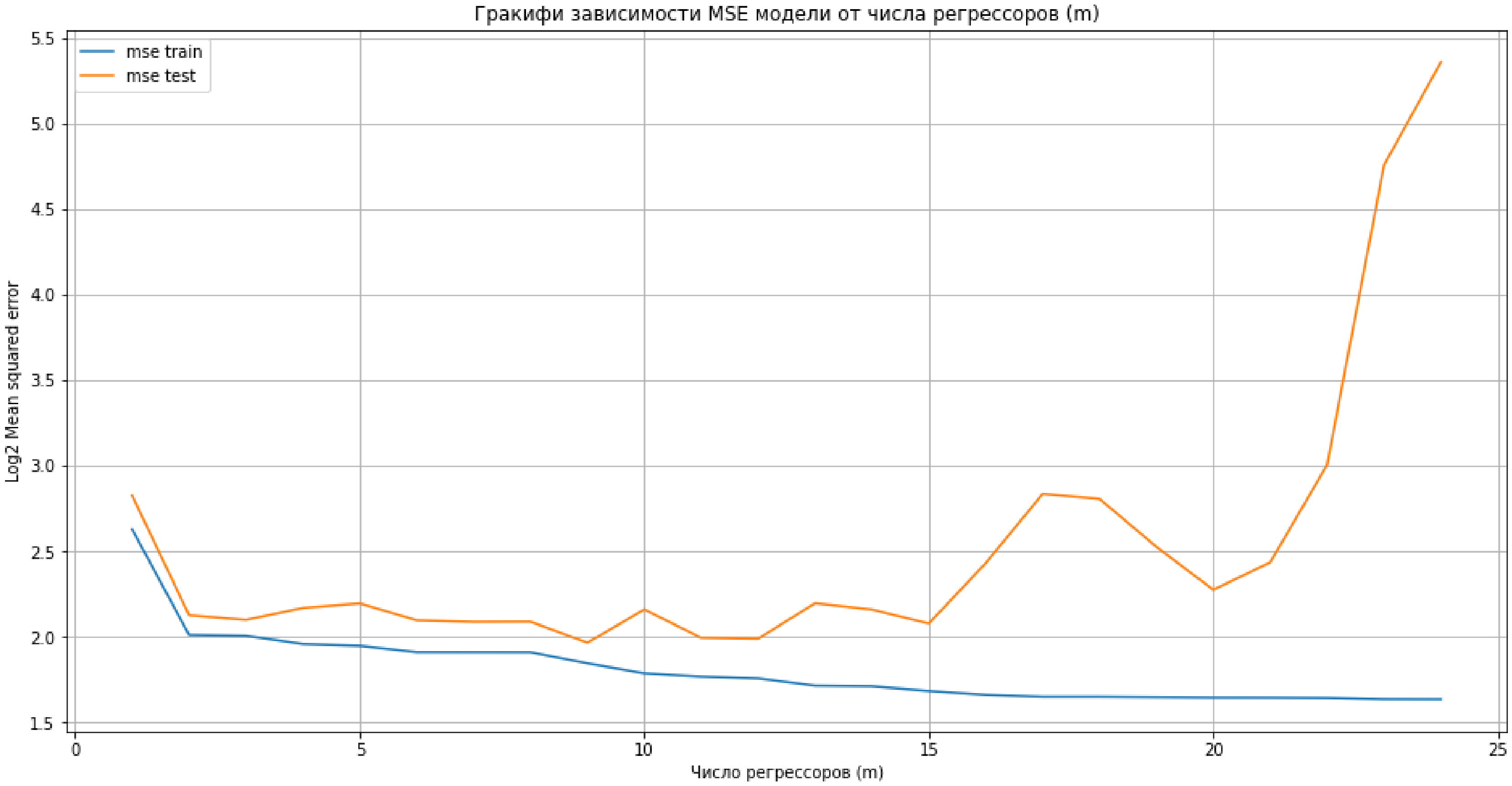
Least-squares criterion: $\mathrm{E}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 \rightarrow min_\beta$

Risk of the model h at given $x \in \mathcal{X}$ (expectation over training samples $\mathcal{D}_T$):

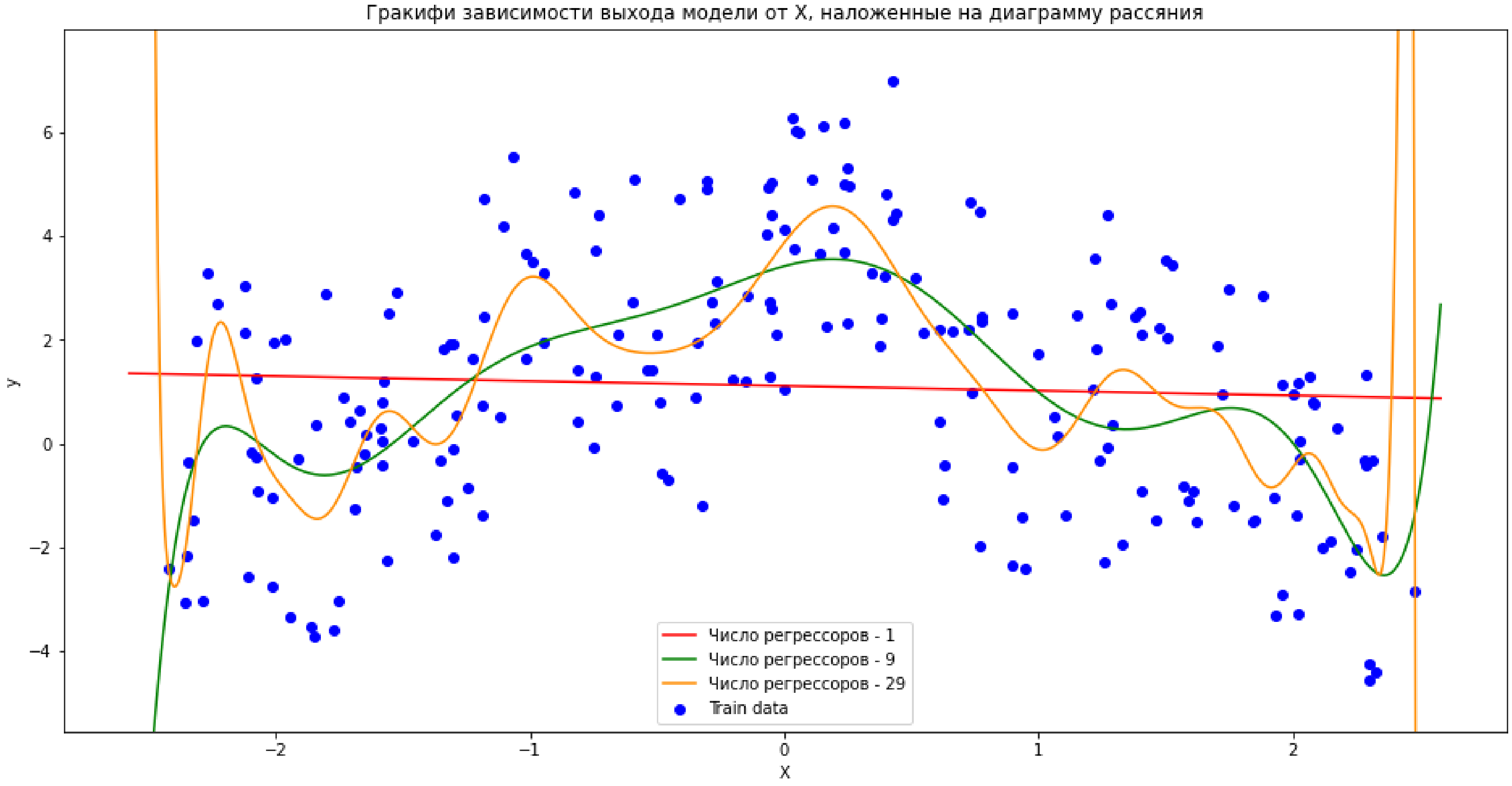
$$\mathrm{R}(h, x) = (M[h(x, \mathcal{D}_T)] - M[Y|_x])^2 + D[h(x, \mathcal{D}_T)] + \sigma_x^2$$

$$\mathrm{R}(h, x) = Bias^2[h] + D[h] + \sigma_x^2$$

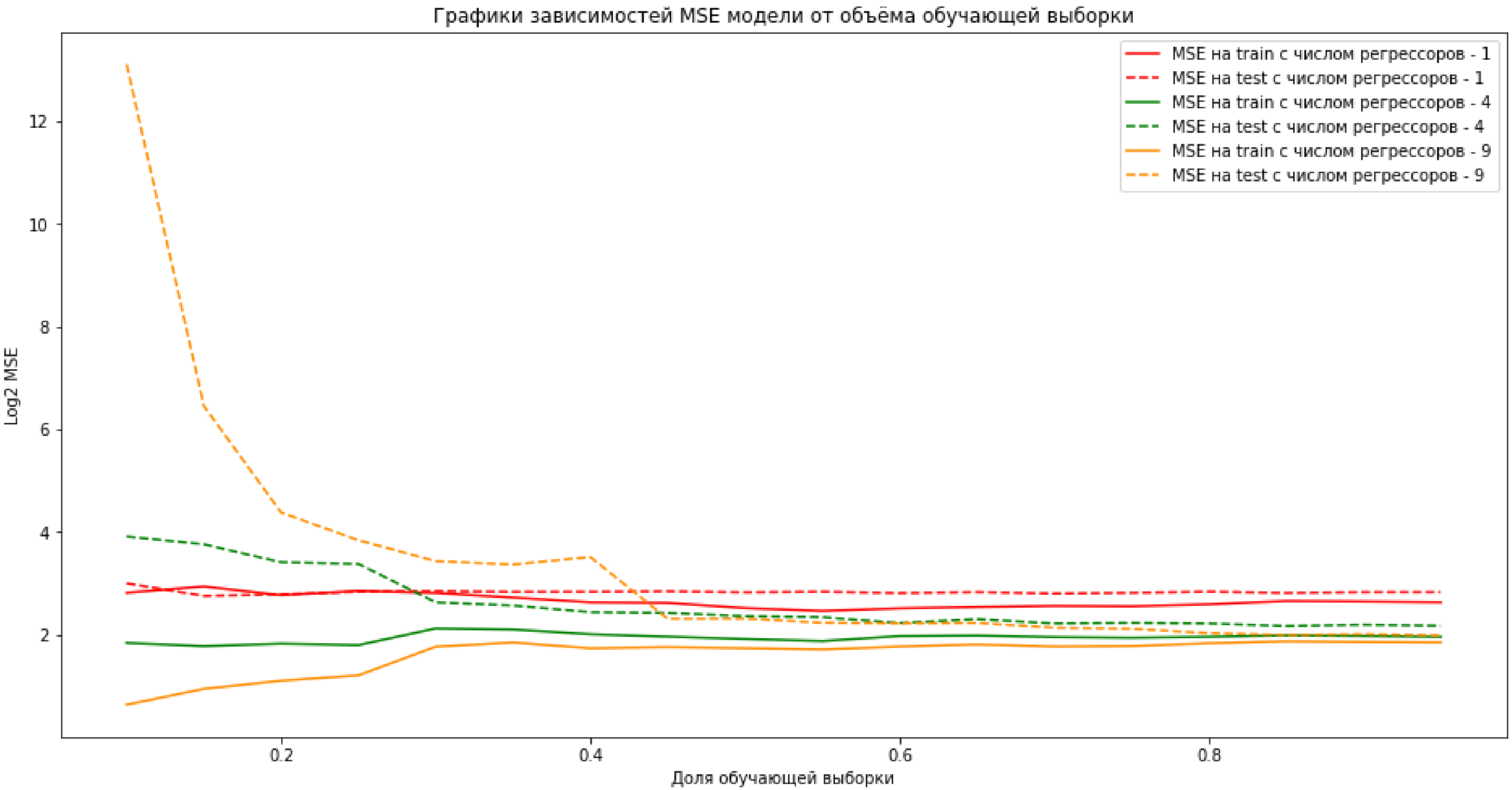
Результаты исследований



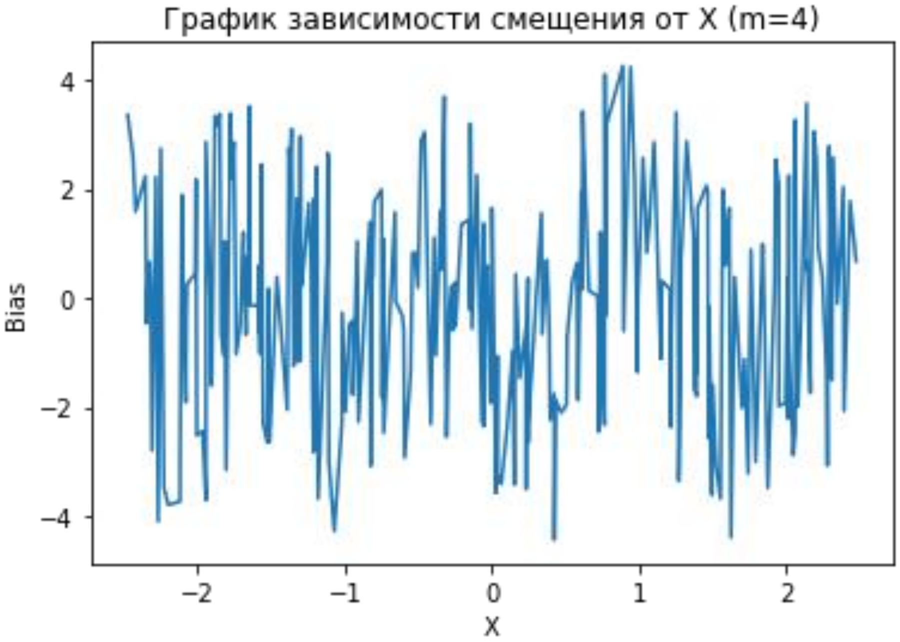
Результаты исследований

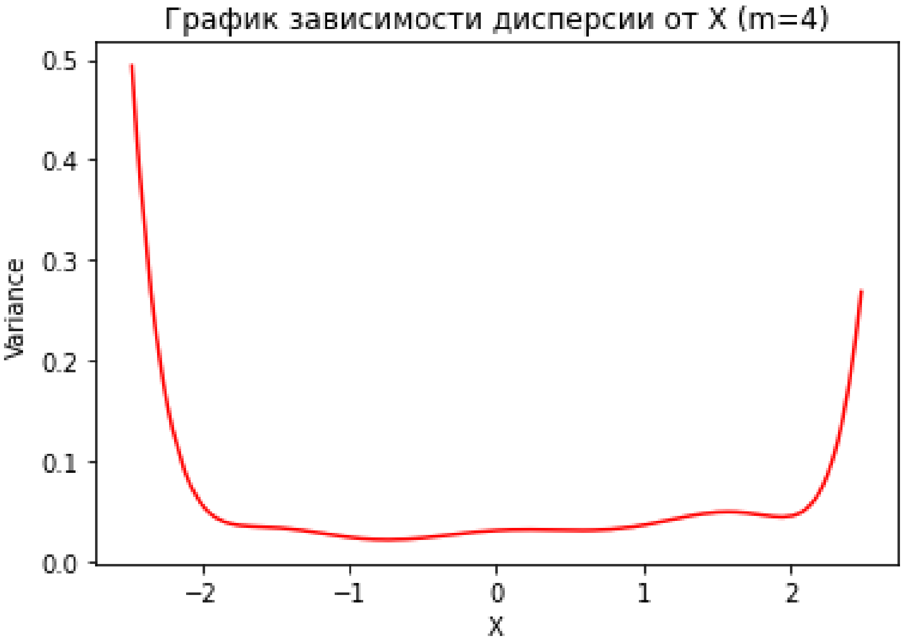


Результаты исследований

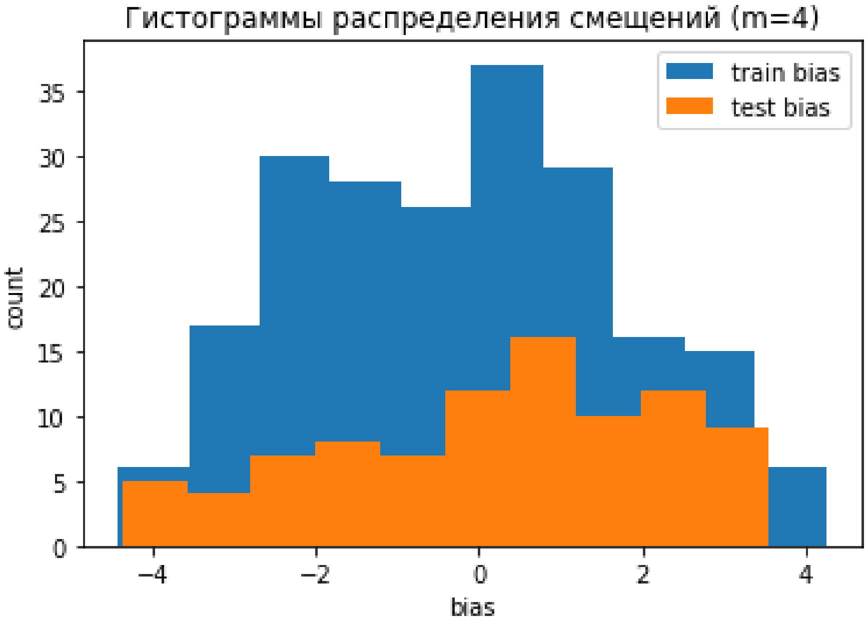


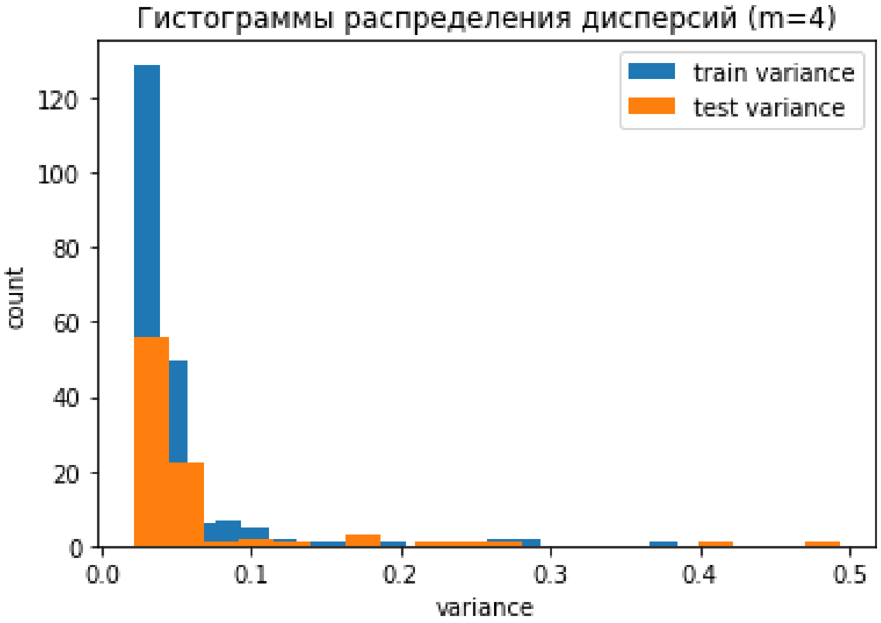
Результаты исследований



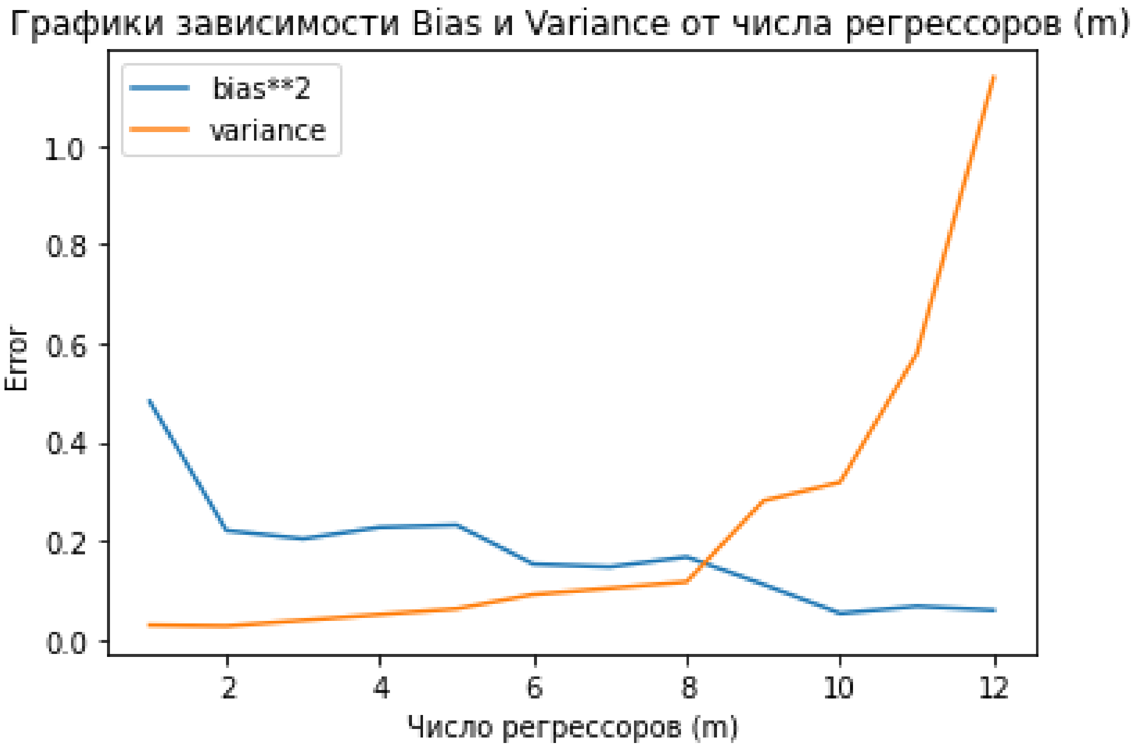


Результаты исследований





Результаты исследований



Выводы

- Высокий *bias* говорит об underfitting;
- Высокий *variance* говорит об overfitting;
- При увеличении числа регрессоров ошибка на обучающей выборке уменьшается;
- При увеличении числа регрессоров *m bias* уменьшается, а *variance* увеличивается.