



Data learning

Курс “Машинное обучение”
Лабораторная работа

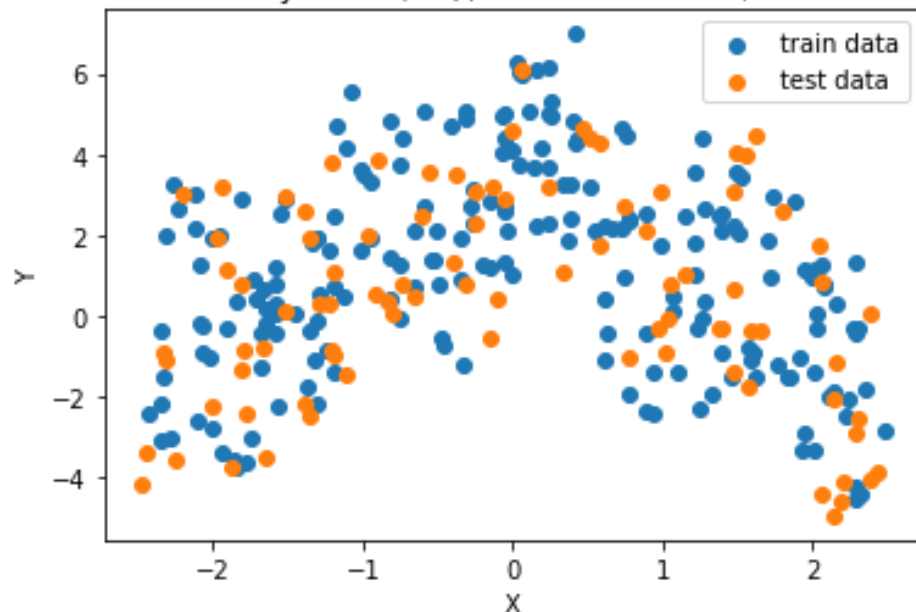


Bias-Variance decomposition

Глушков А.Е., М21-524
Вариант 1-01

Исходные данные

Визуализация данных Holdout(70/30)



	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:

$$\varphi(x) = M[Y|x] = \beta_0 + \beta_1 x$$

Regression models class:

$$h(x) = \beta_0 + \sum_{i=1}^m \beta_i x^i$$

Least-squares criterion:

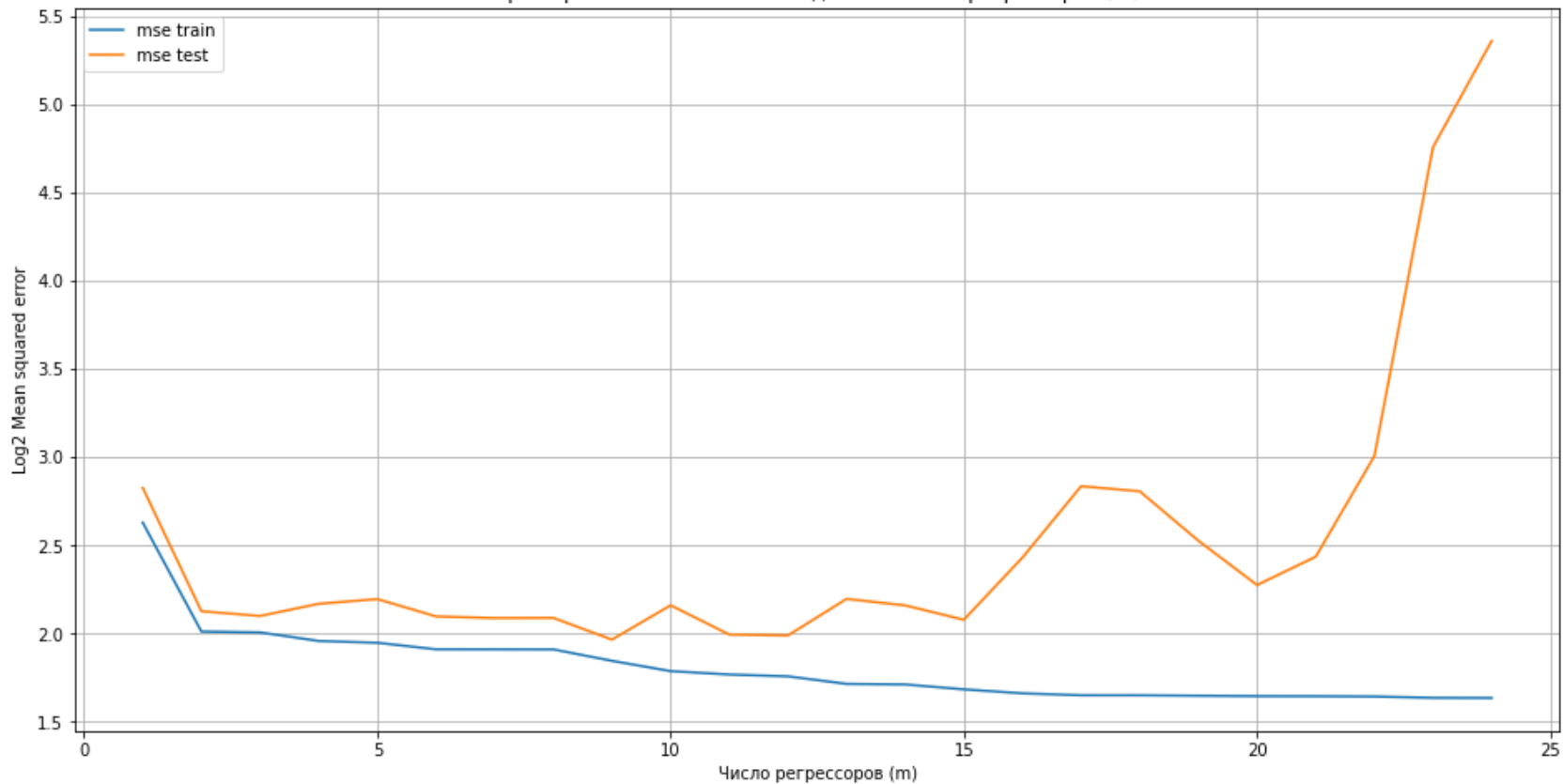
$$E(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i))^2 = \frac{1}{n} \sum_{i=1}^n (y_i - x_i \beta)^2 \rightarrow \min_{\beta}$$

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

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

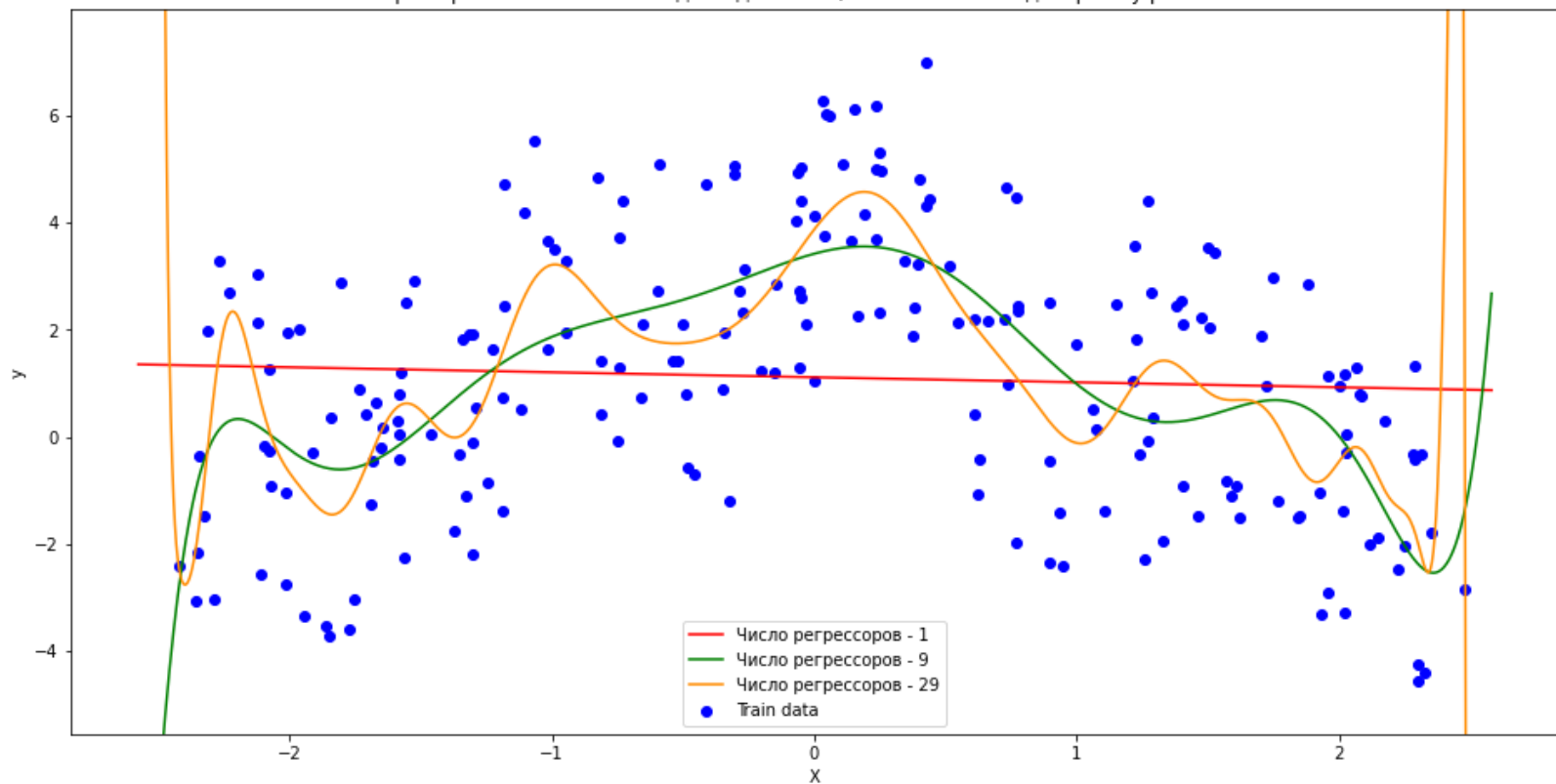
Результаты исследований

Графики зависимости MSE модели от числа регрессоров (m)



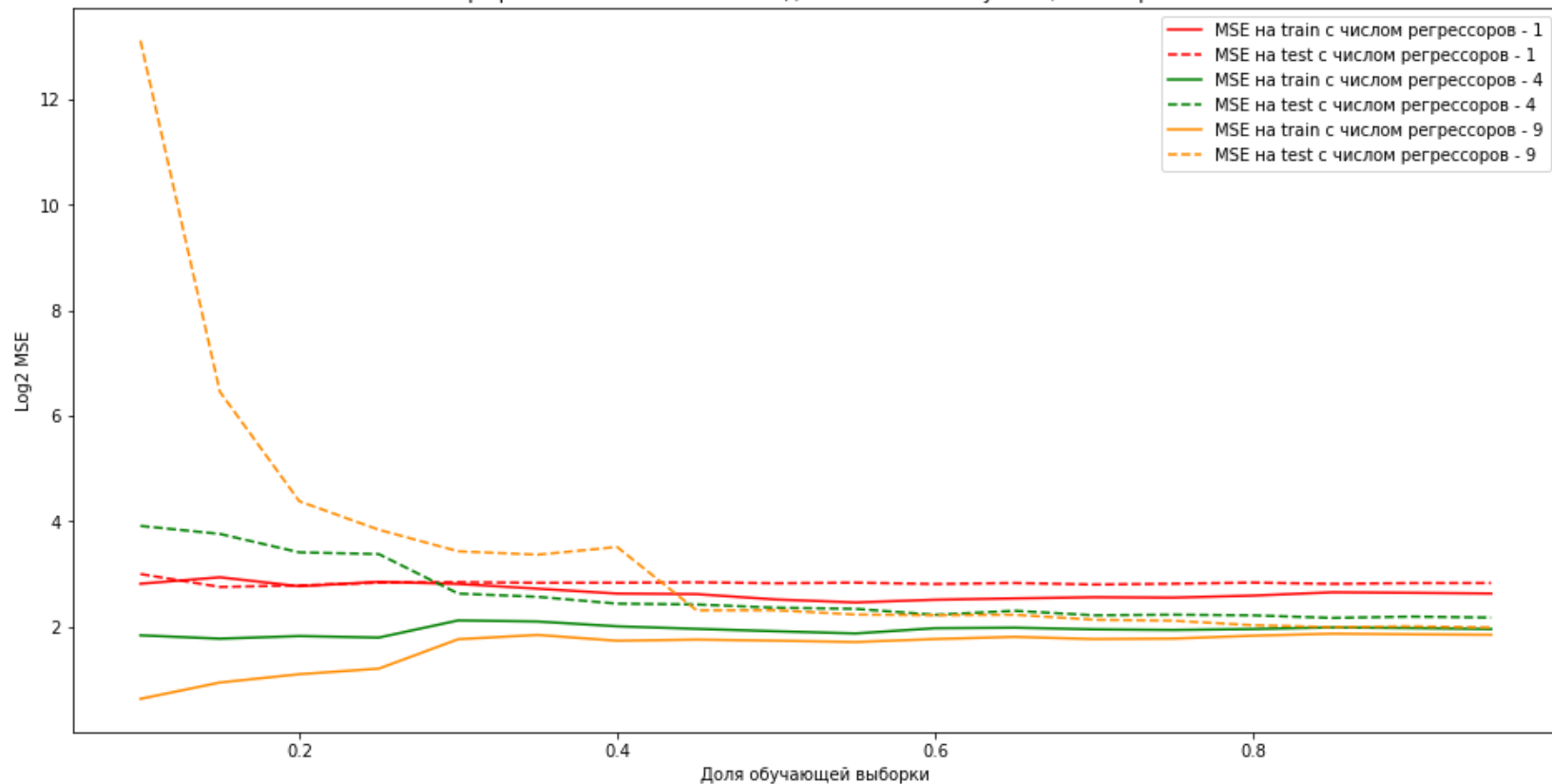
Результаты исследований

Графики зависимости выхода модели от X , наложенные на диаграмму рассеяния



Результаты исследований

Графики зависимостей MSE модели от объёма обучающей выборки



Результаты исследований

График зависимости смещения от X ($m=4$)

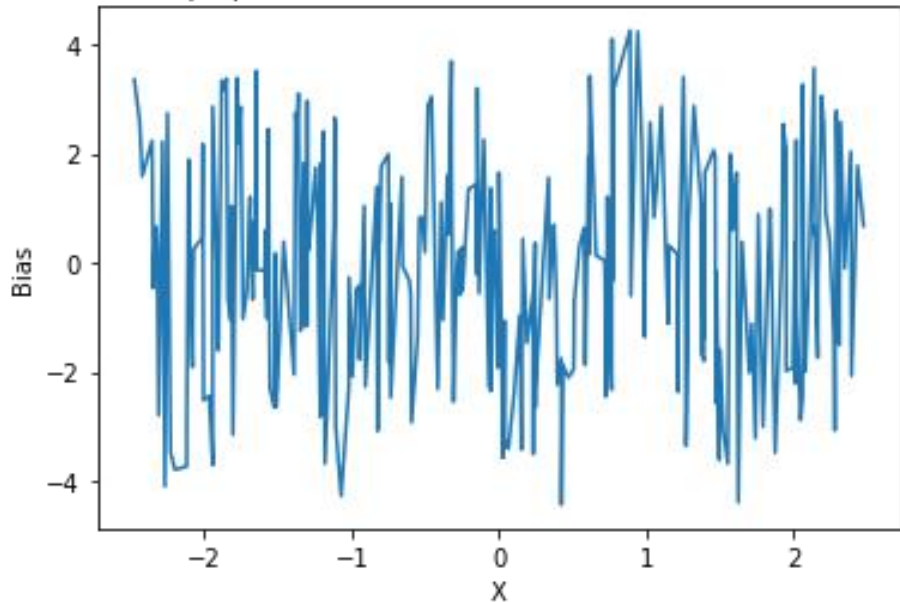
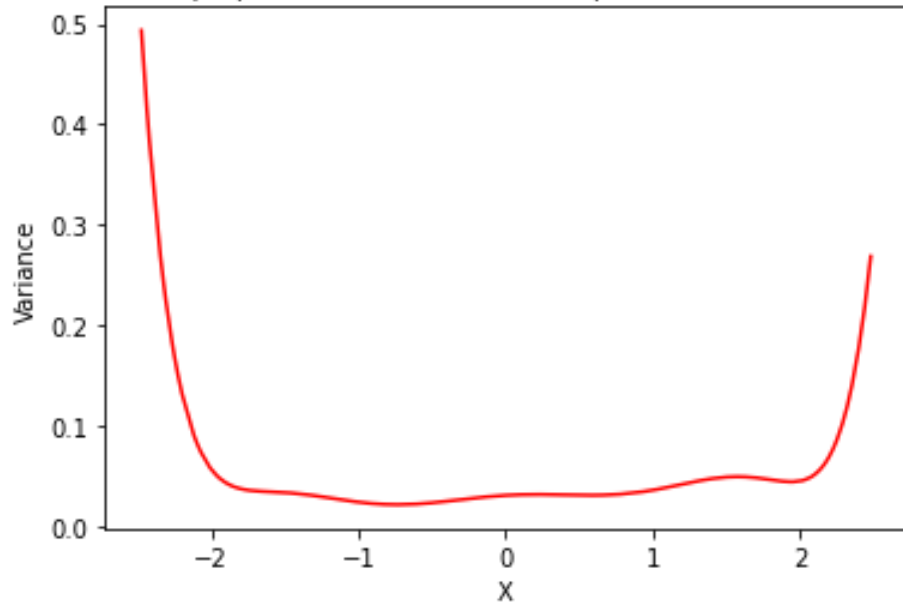
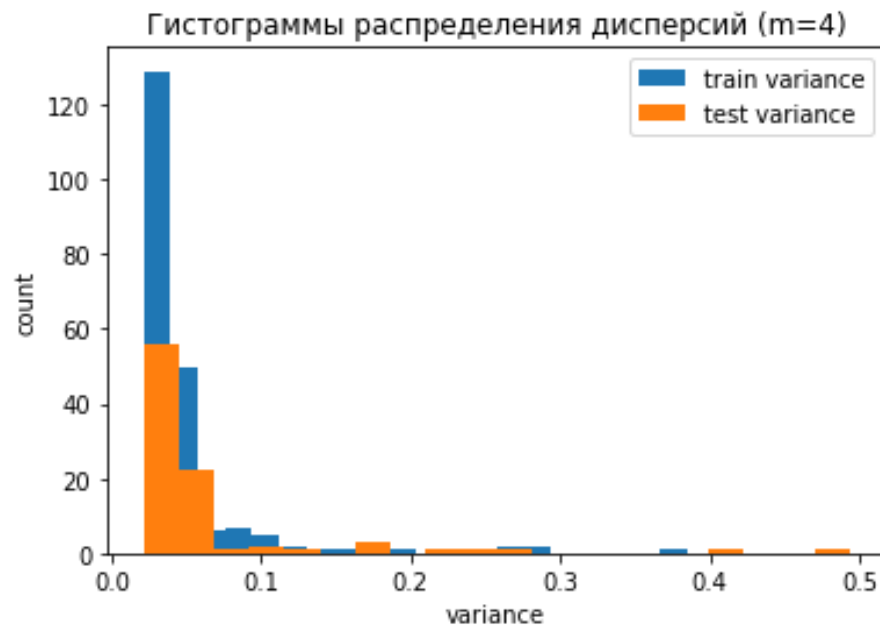
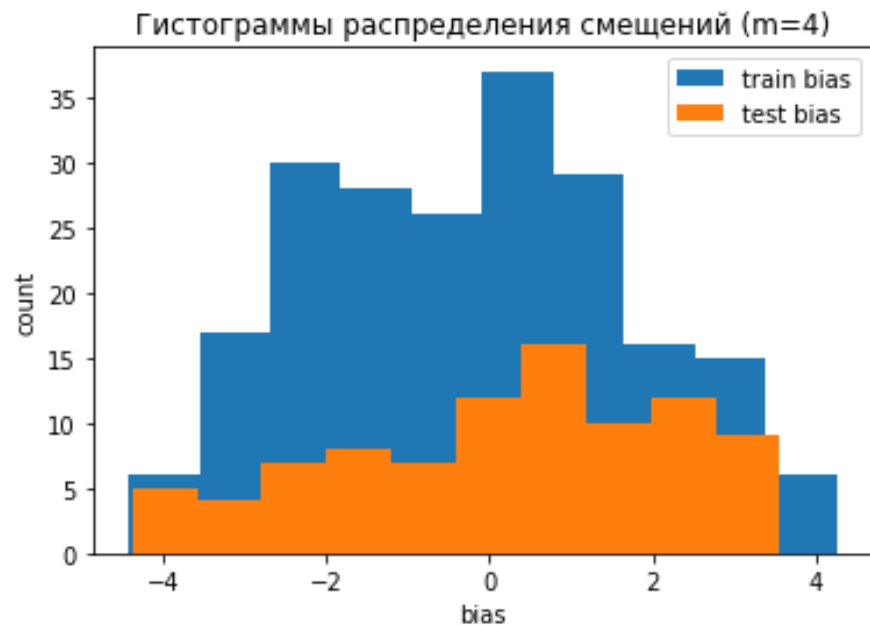


График зависимости дисперсии от X ($m=4$)

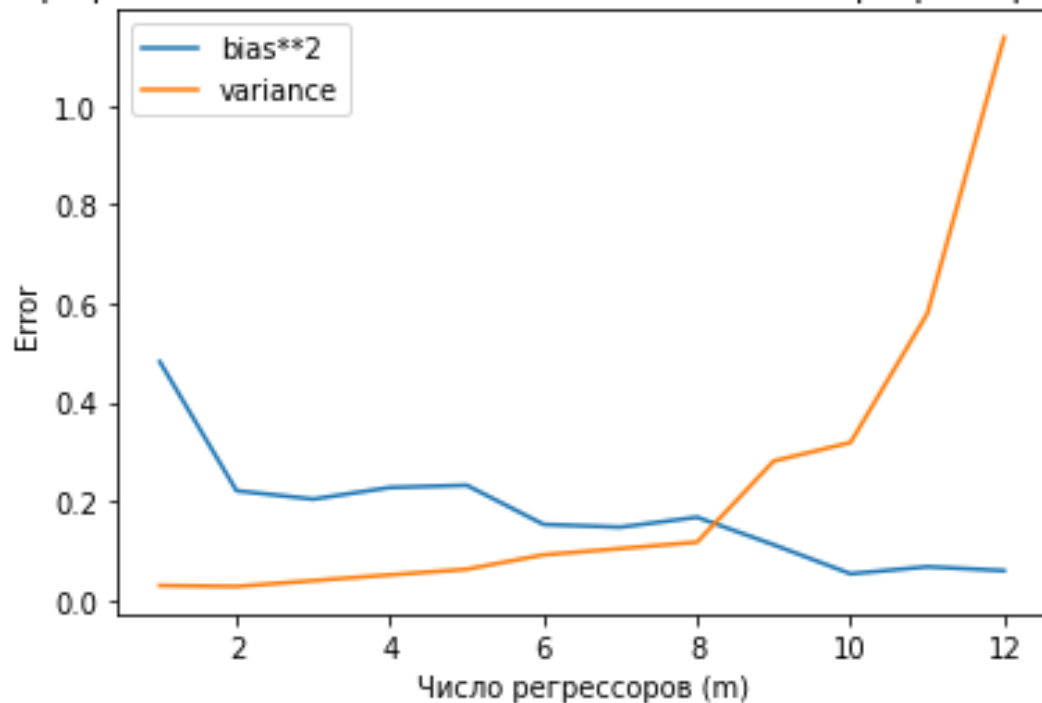


Результаты исследований



Результаты исследований

Графики зависимости Bias и Variance от числа регрессоров (m)



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.

