Проектный практикум «Модель прогнозирования времени до землетрясения по данным лабораторных симуляций»

Студенты:

- Токарева Анастасия
- Великоречанин Игорь

Направление подготовки: «Науки о данных», МФТИ

Цель

- **Построить модель предсказания времени до землетрясения**

Дано

- **Выборка лабораторных экспериментов.**

Пара значений:

1. **Акустические данные – X**
2. **Время до землетрясения - Y**

МФТИ

Преобразование

- **Проблемы**

  1. Наблюдения не происходят равномерно по времени
  2. Наблюдения слишком частые

- **Обработка**

  1. Преобразование данных оконными функциями
  2. Шаг вычисления 10 мс, ширина окна 20 мс
  3. Рассматриваемые статистики:

- $F_A(W) = \frac{1}{|W|}\sum_{i \in W} x_i$
- $F_B(W) = \frac{1}{|W|}\sum_{i \in W}(x_i - \frac{1}{|W|}\sum_{j \in W} x_j)^2$
- $F_C(W) = \frac{1}{|W|}\sum_{i \in W}\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$

- **Пример данных**

acoustic_data	time_to_failure
12	1.4690999832
6	1.4690999821
8	1.469099981
5	1.4690999799
8	1.4690999777

mean (acoustic_data)	mean (time_to_failure)
4.519467573700124	5.678291712978854

МФТИ

Данные с лагами

- **Обозначим**

$$D_p = \begin{pmatrix} A_1 & B_1 & C_1 & \dots & A_{1+p} & B_{1+p} & C_{1+p} \\ A_1 & B_1 & C_1 & \dots & A_{1+p} & B_{1+p} & C_{1+p} \\ - & - & - & - & - & - & - \\ A_{m-p} & B_{m-p} & C_{m-p} & \dots & A_m & B_m & C_m \end{pmatrix} \quad Q_p = \begin{pmatrix} Y_{1+p} \\ Y_{2+p} \\ - \\ Y_m \end{pmatrix}$$

- **Где A_i, B_i, C_i – значения оконных функций в соответствии с моментом времени**

МФТИ

Регрессионные модели

- **Входные данные:**

  Значения окон с p лагами

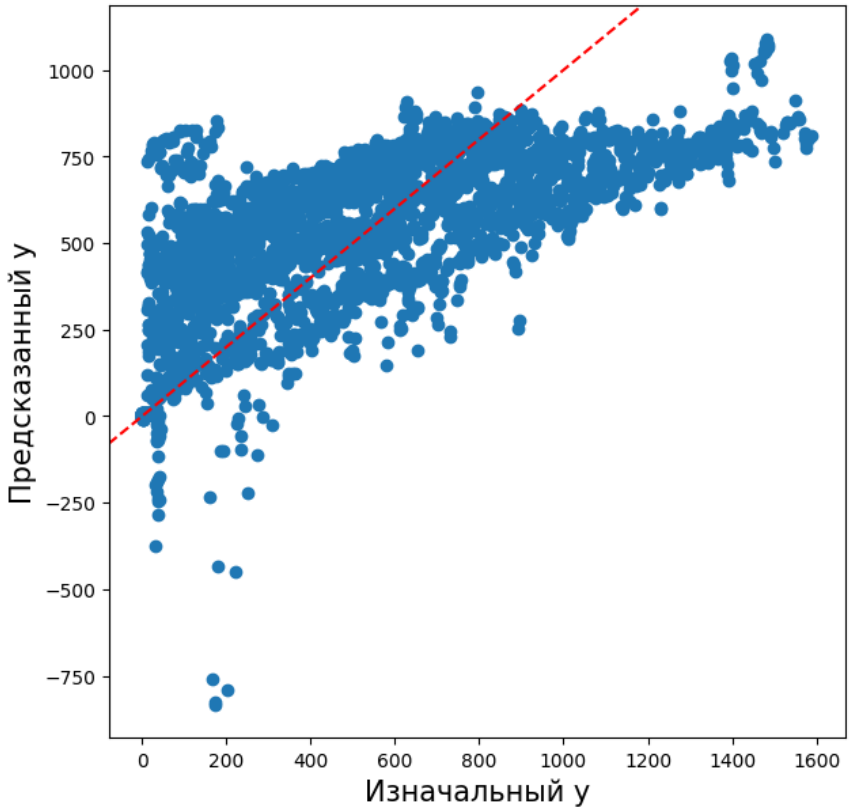
- **Рассматриваемые модели:**

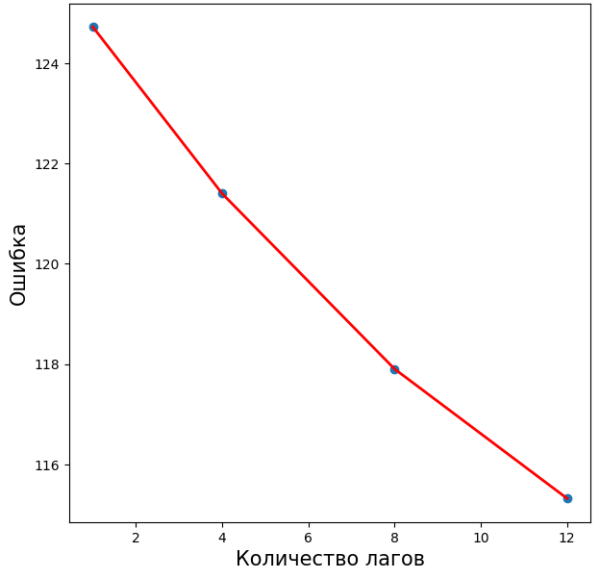
  1. **Линейная регрессия**
  2. **LASSO**
  3. **XGBoost**

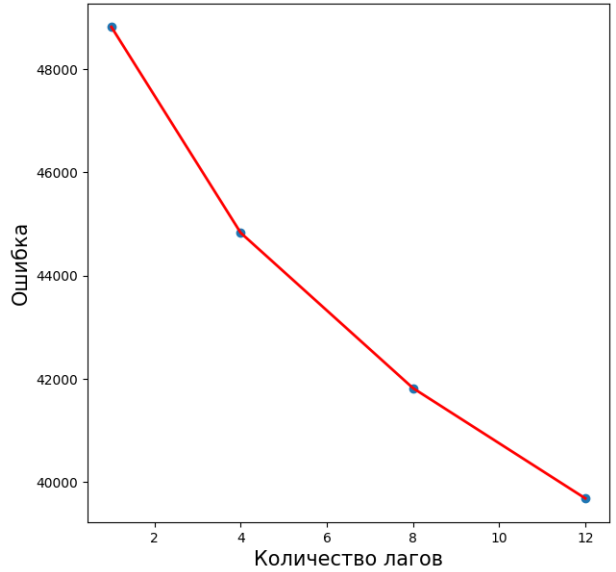
- **Оценка точности**

  1. **Коэффициент детерминации R^2**
  2. **MAE**
  3. **MSE**

График сравнения предсказанных и истинных значений отклика для модели линейной регрессии с количеством лагов 1 и 12. График зависимости R2, MAE и MSE от кол-ва лагов.







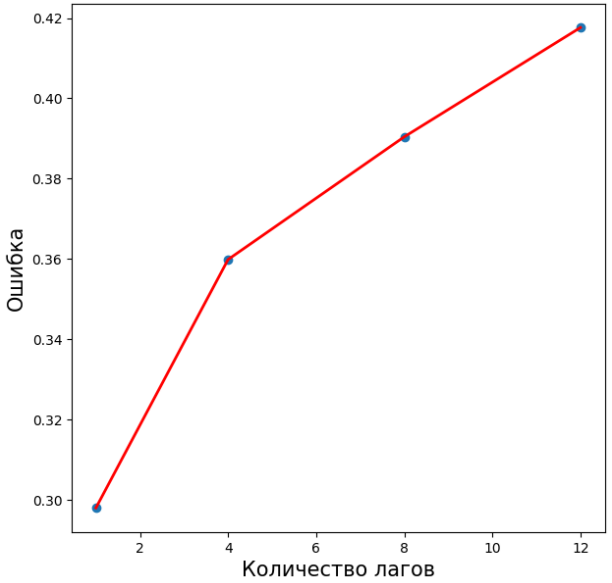
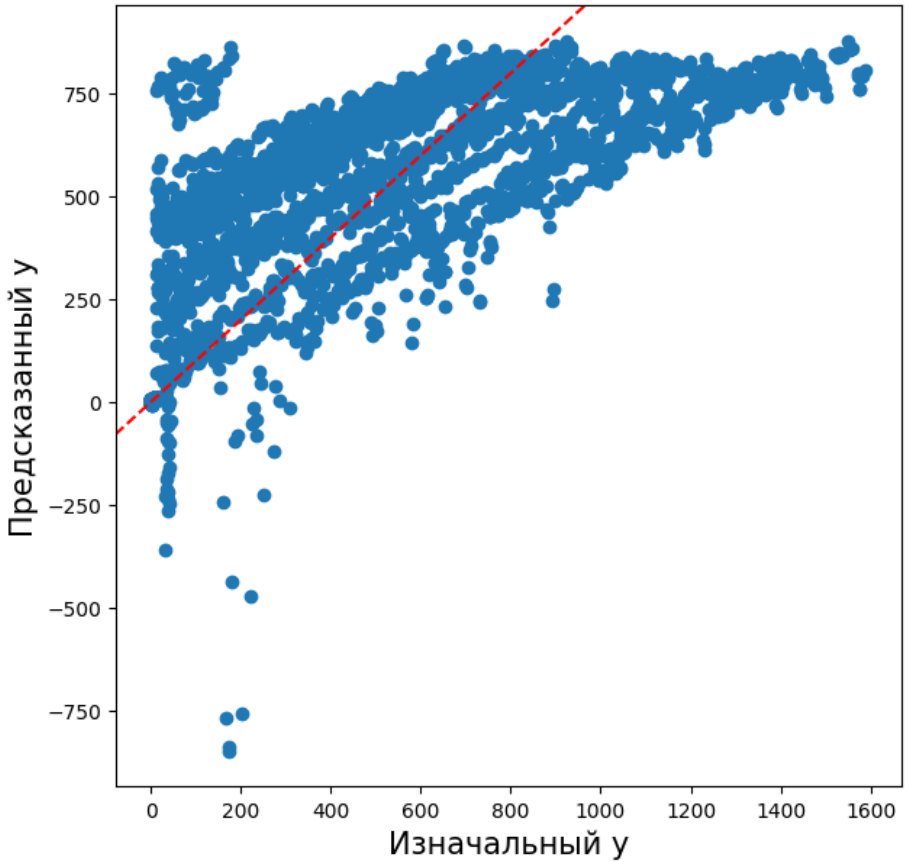
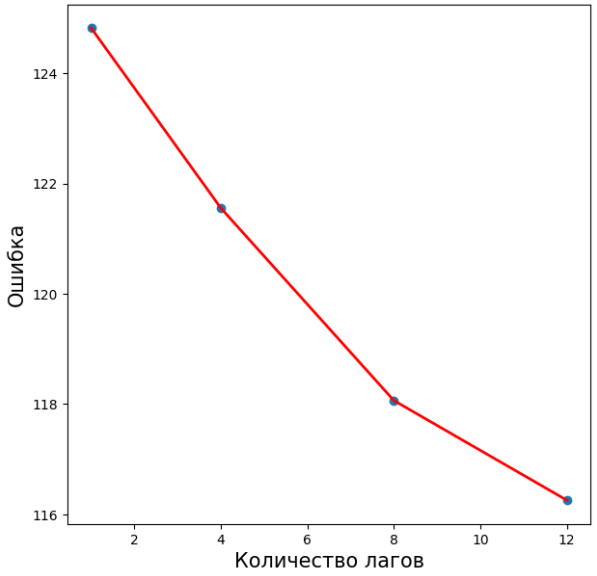
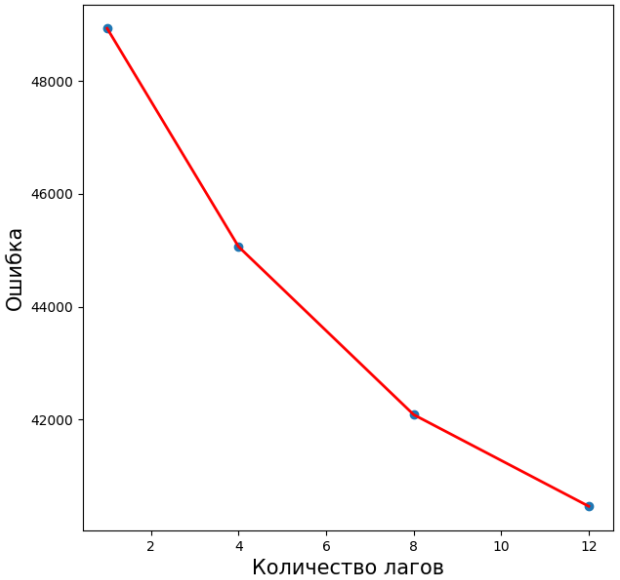


График сравнения предсказанных и истинных значений отклика для модели LASSO с количеством лагов 12. График зависимости R2, MAE и MSE от кол-ва лагов.







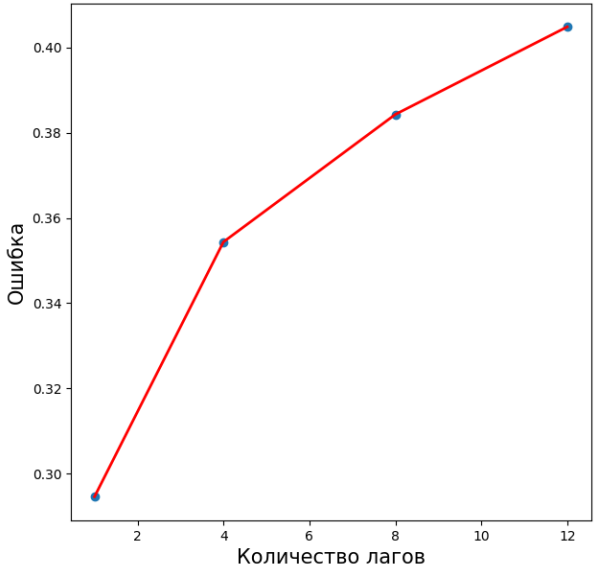
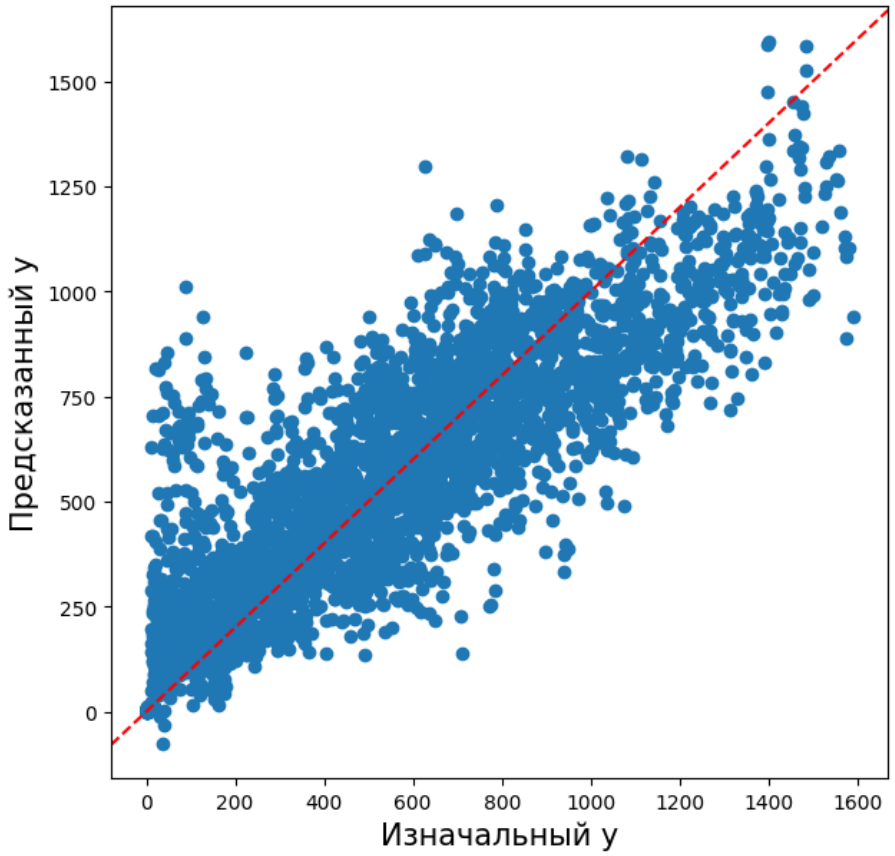
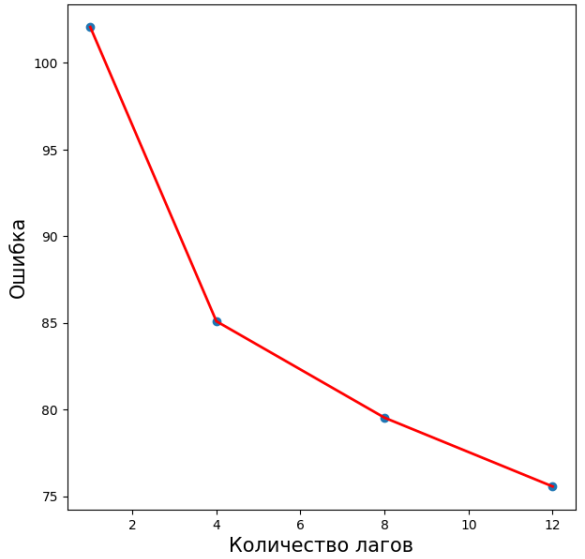
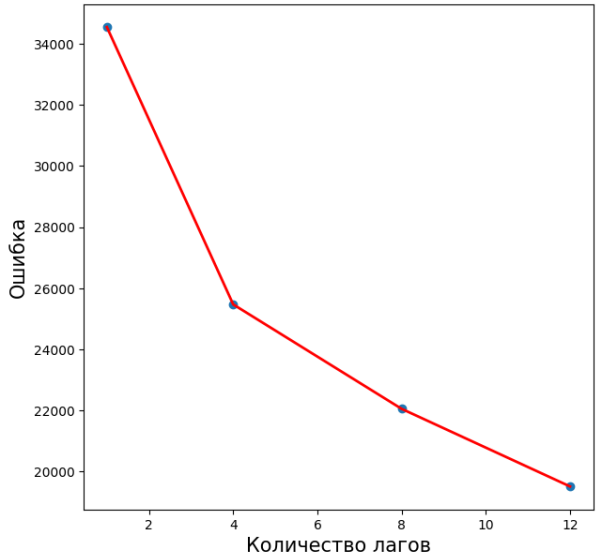
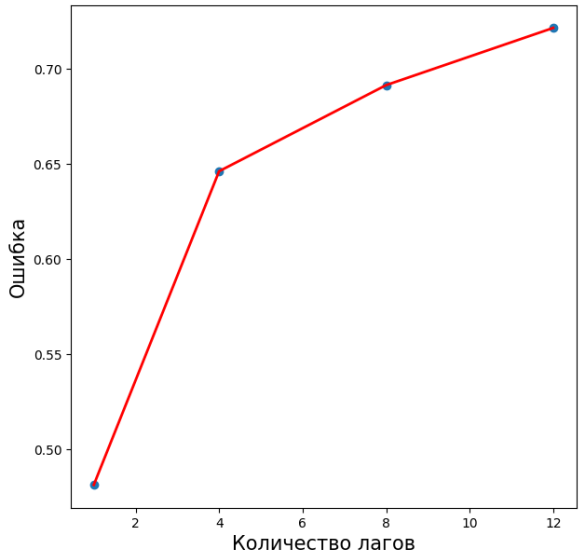


График сравнения предсказанных и истинных значений отклика для модели XGBoost с количеством лагов 12. График зависимости R2, MAE и MSE от кол-ва лагов.









Результаты

	D_1	D_2	D_8	D_{12}
Лин. регрес	0.2979	0.3598	0.3904	0.4177
LASSO	0.2947	0.3544	0.3843	0.4048
XGBoost	0.4811	0.6461	0.6915	0.7215
Cross_val_xg	0.4955	0.6656	0.7062	0.7241

Выводы

- При использовании модели XGBoost с p = 12 достигается самая высокая точность $R^2 = 0.7215$

Заключение

- Мы преобразовали начальные данные перед обучением с помощью оконных функций и трех статистик
- Рассмотрели три регрессионные модели с различными лагами
- Применили кросс-валидацию к модели XGBoost и построили графики зависимостей
- Достигли точности $R^2 = 0.7241$ при кросс-валидации

МФТИ