Собеседование на специализацию «Интеллектуальный анализ данных»

Киселев Никита Б05-002

18 апреля 2022 г.

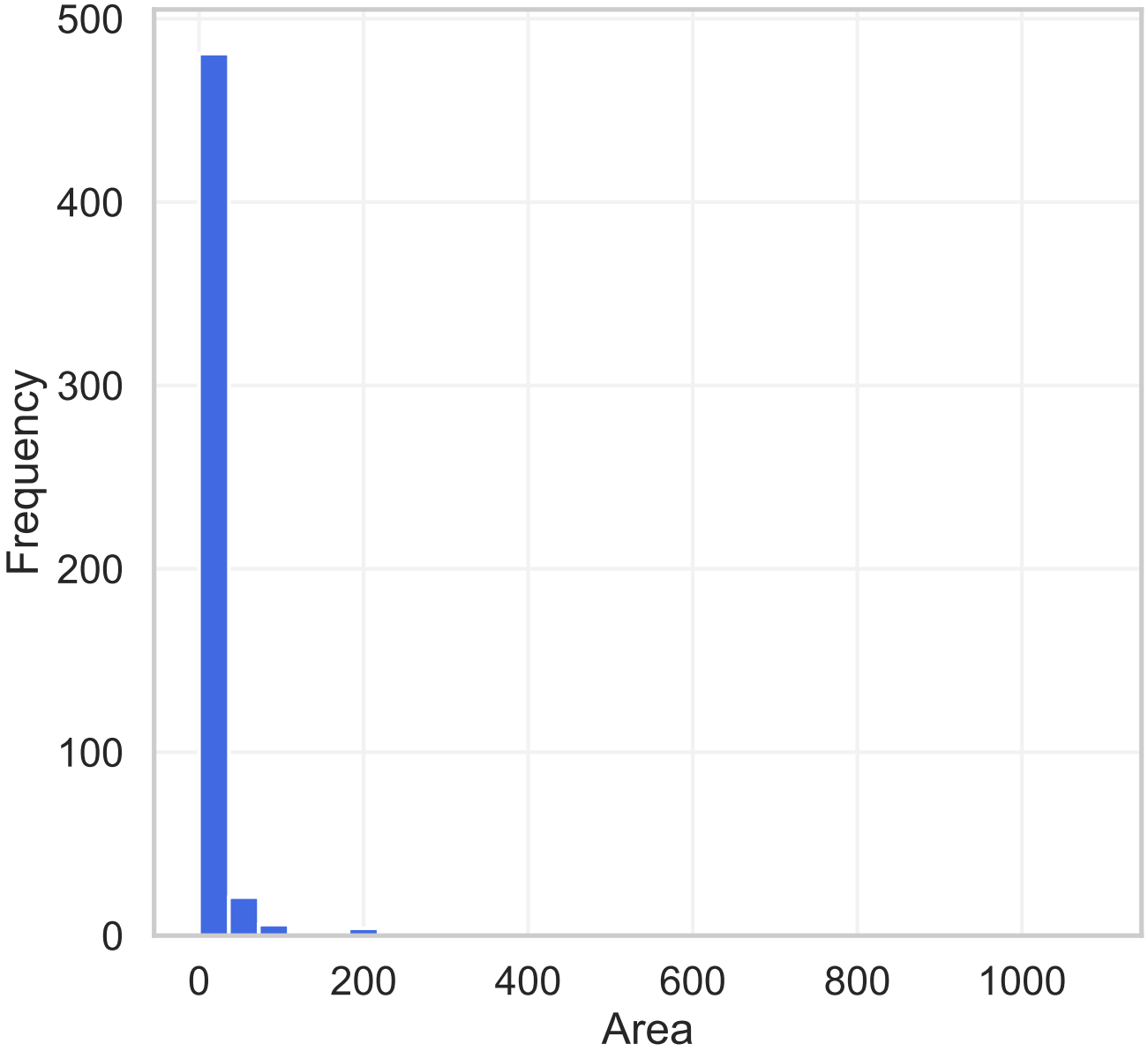
Московский физико-технический институт
(национальный исследовательский университет)

Тестовая задача

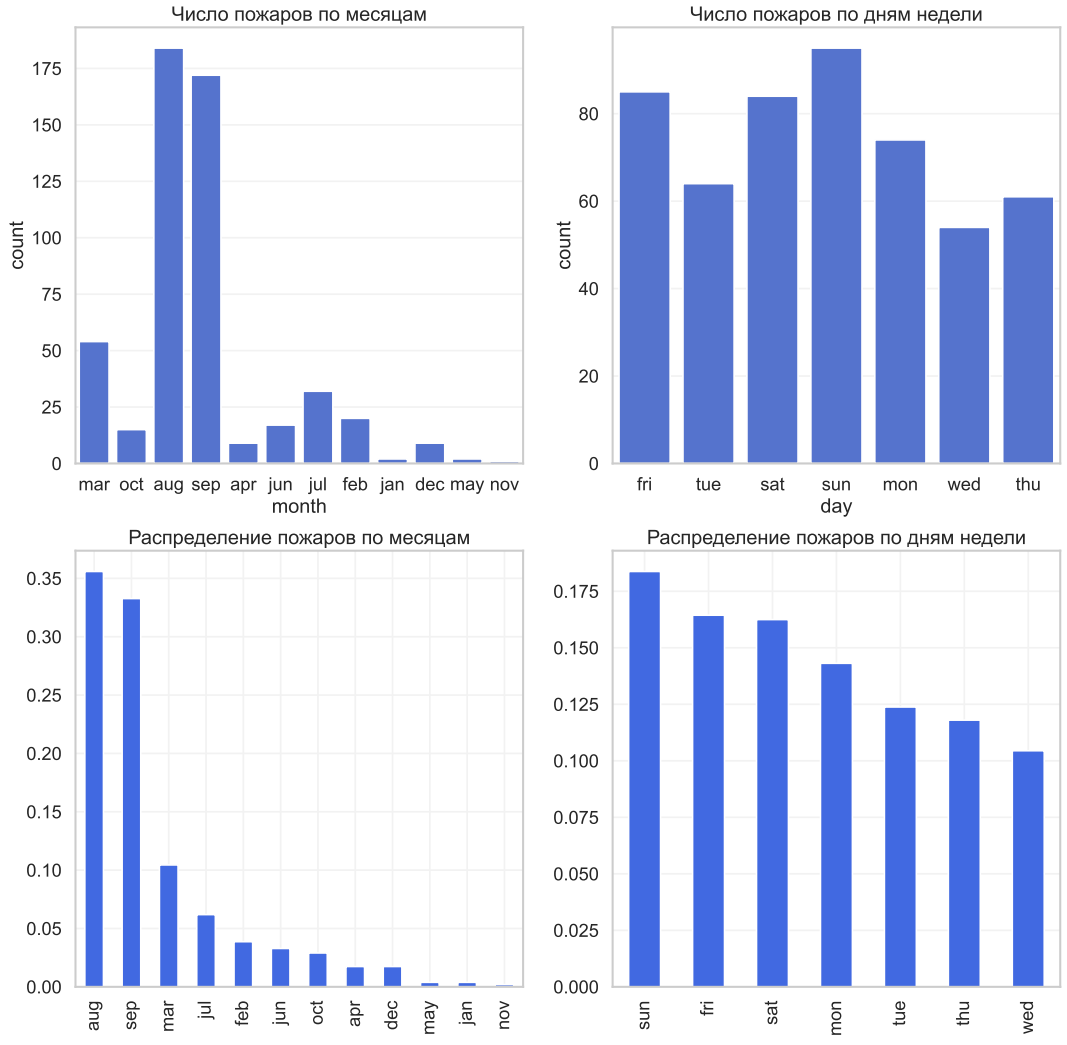
Задача 21

Предсказание площади лесных пожаров. На основе погодных измерений необходимо предсказать объем выгоревших лесных массивов на севере Португалии. Выборка состоит из 13 признаков и 517 объектов. Для решения задачи предлагается использовать метод наименьших квадратов с регуляризацией. Нарисовать график весов признаков и общей ошибки на кросс-валидации при изменении параметра регуляризации. Какие признаки наиболее важны для нашей задачи? Что изменится, если предварительно все признаки стандартизовать?

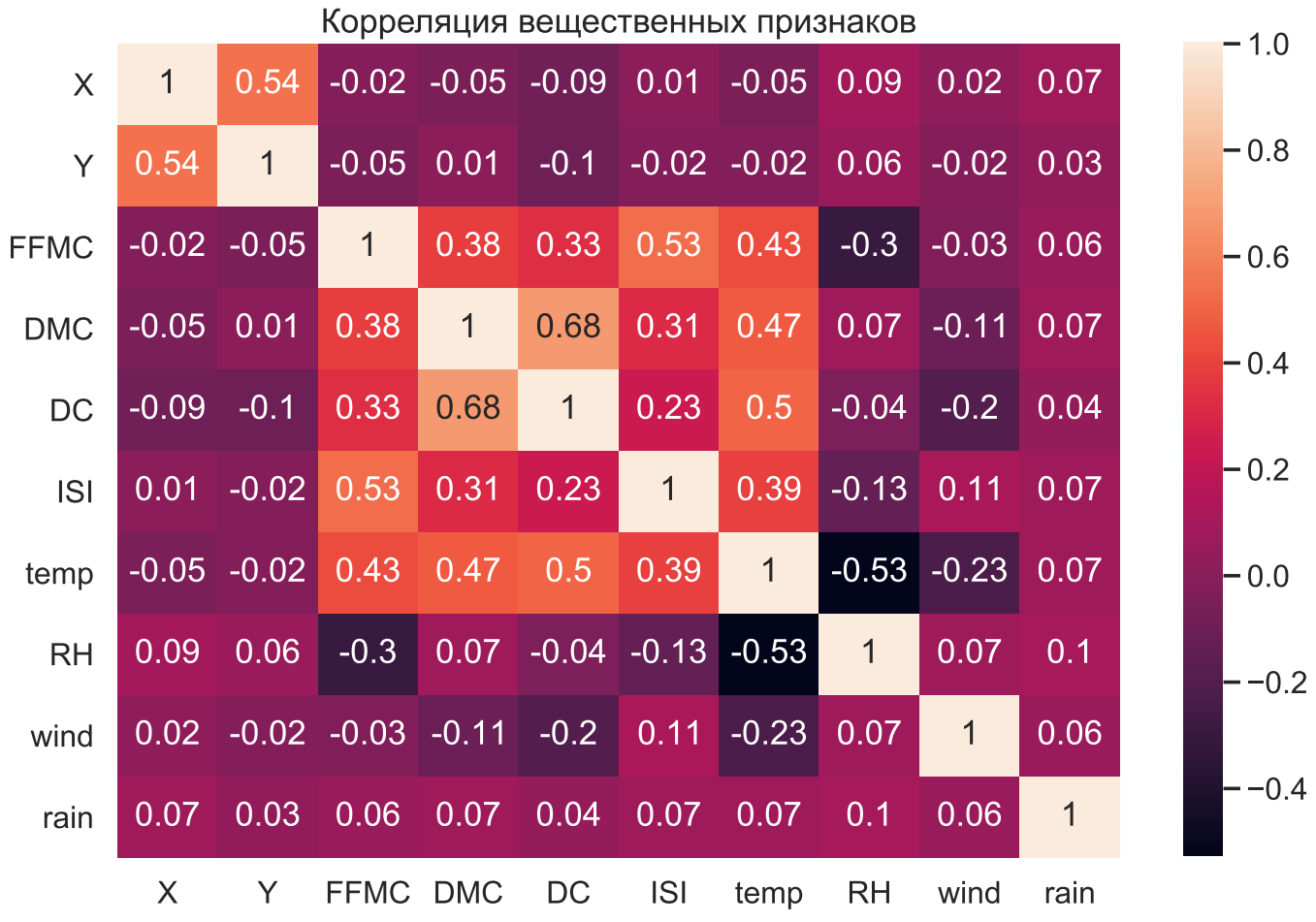
Распределение ответов



Распределение номинальных признаков



Корреляция количественных признаков



Линейная регрессия

- Множество объектов $\mathbb{X} = \mathbb{R}^n$
- Объекту $\boldsymbol{x} \in \mathbb{X}$ соответствует признаковое описание $\boldsymbol{x} = (f_1(x), \ldots, f_n(x))$, где $f_j : \mathbb{X} \to D_j$
- Множество ответов $\mathbb{Y} = \mathbb{R}$
- Выборка $\mathbb{D} = \{(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i \in \mathbb{X}, y_i \in \mathbb{Y}, i = 1, \ldots, m\}$
- Матрица объекты-признаки $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)^T$, вектор ответов $\boldsymbol{y} \in \mathbb{Y}^m$
- Вектор параметров модели $\boldsymbol{w} = (w_1, \ldots, w_n)^T$
- Ставится задача минимизации ошибки алгоритма $Q(\boldsymbol{w}, X) = \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 \to \min\limits_{\boldsymbol{w}}$

Метод наименьших квадратов

$$Q(\boldsymbol{w}, X) = \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 = (X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y}) \to \min_{\boldsymbol{w}}$$

Приравняем к нулю производную по вектору $\boldsymbol{w}$:

$$\nabla_{\boldsymbol{w}} Q(\boldsymbol{w}, X) = \nabla_{\boldsymbol{w}}(-\boldsymbol{y}^T X\boldsymbol{w} + \boldsymbol{w}^T X^T X\boldsymbol{w} + \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T X^T\boldsymbol{y}) =$$
$$= -X^T\boldsymbol{y} + (X^T X + X^T X)\boldsymbol{w} + 0 - X^T\boldsymbol{y} = 0$$

$$X^T X\boldsymbol{w} = X^T\boldsymbol{y}$$

$$\boxed{\boldsymbol{w}^* = (X^T X)^{-1} X^T\boldsymbol{y}}$$

L_2 регуляризация

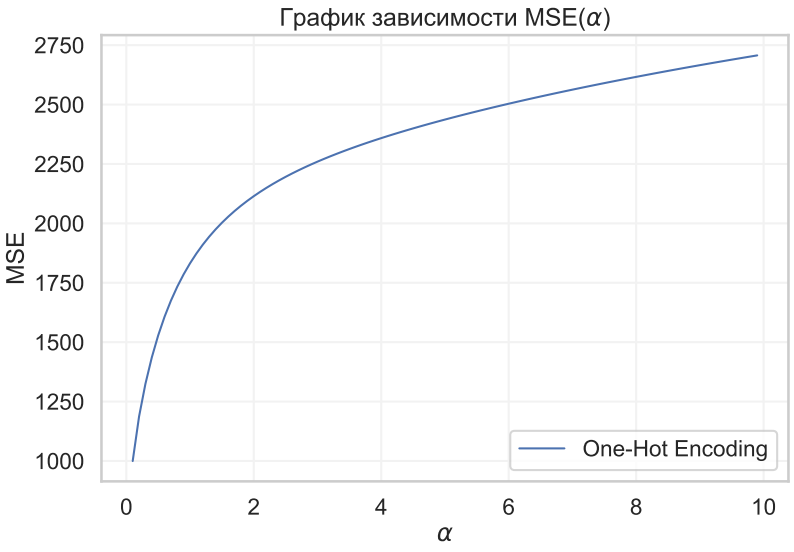
Могут возникнуть проблемы мультиколлинеарности в случае, если матрица $X^T X$ плохо обусловлена. Один из способов решения — добавление к этой матрице диагональной:

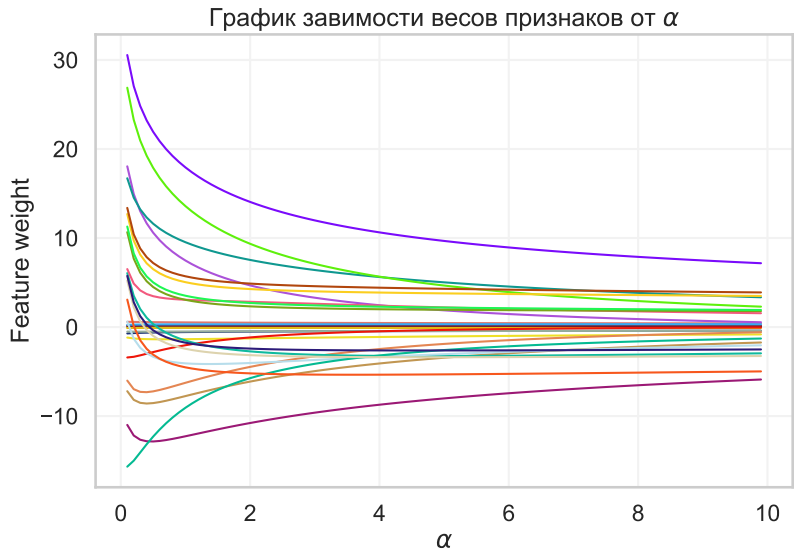
$$\boxed{\boldsymbol{w}^* = (X^T X + \alpha E_n)^{-1} X^T \boldsymbol{y}}$$

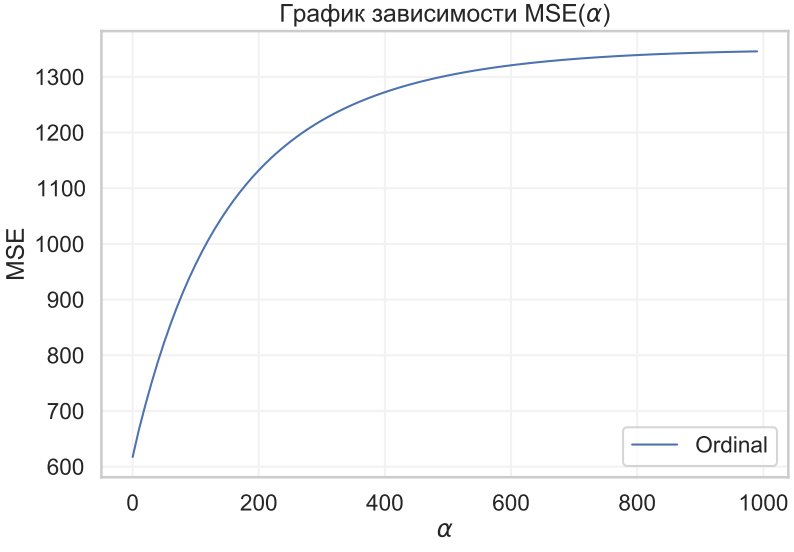
При этом значении вектора $\boldsymbol{w}$ достигается минимум функционала ошибки

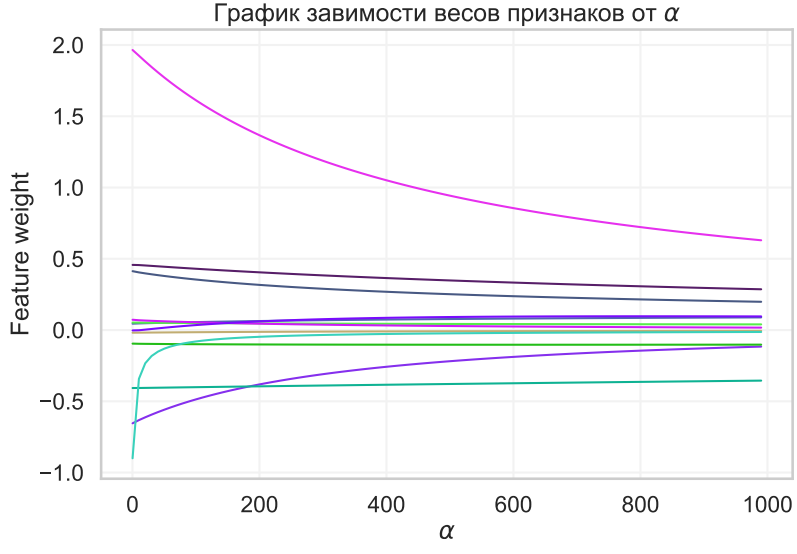
$$Q(\boldsymbol{w}, X, \alpha) = \|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \alpha\|\boldsymbol{w}\|_2^2$$

Изменение параметра α









Стандартизация

При стандартизации происходит преобразование признаков:

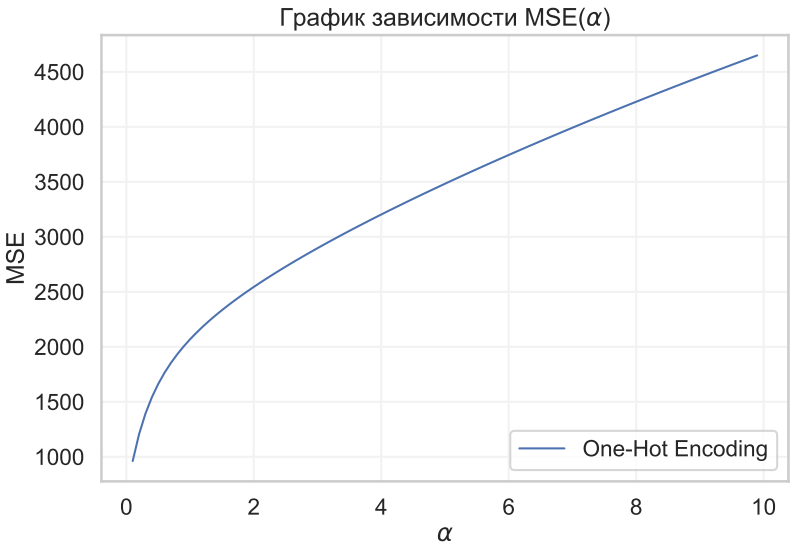
$$\hat{f}_j(\boldsymbol{x}_i) = \frac{f_j(\boldsymbol{x}_i) - \bar{f}_j}{S_j},$$

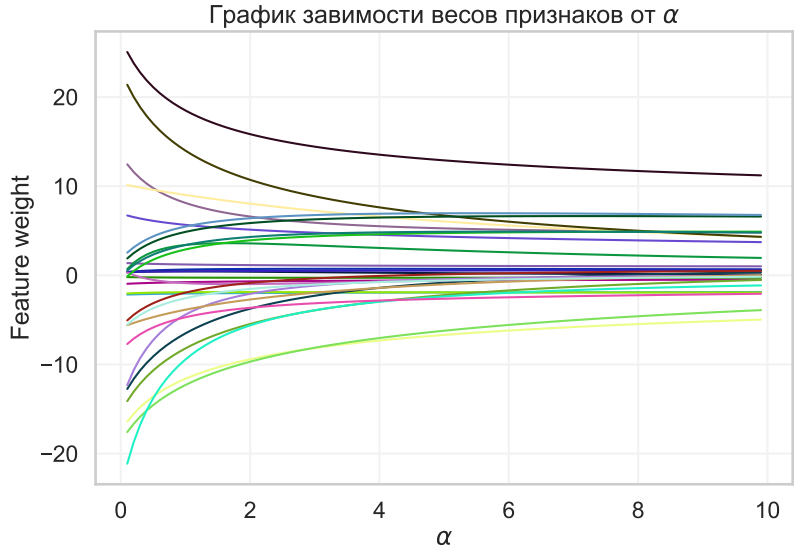
где

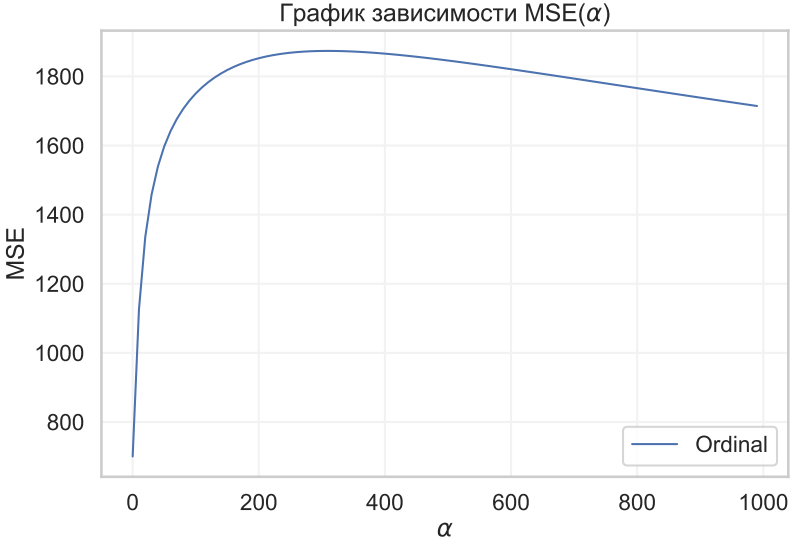
$$\bar{f}_j = \frac{1}{m}\sum_{i=1}^{m} f_j(\boldsymbol{x}_i) \text{ — выборочное среднее,}$$

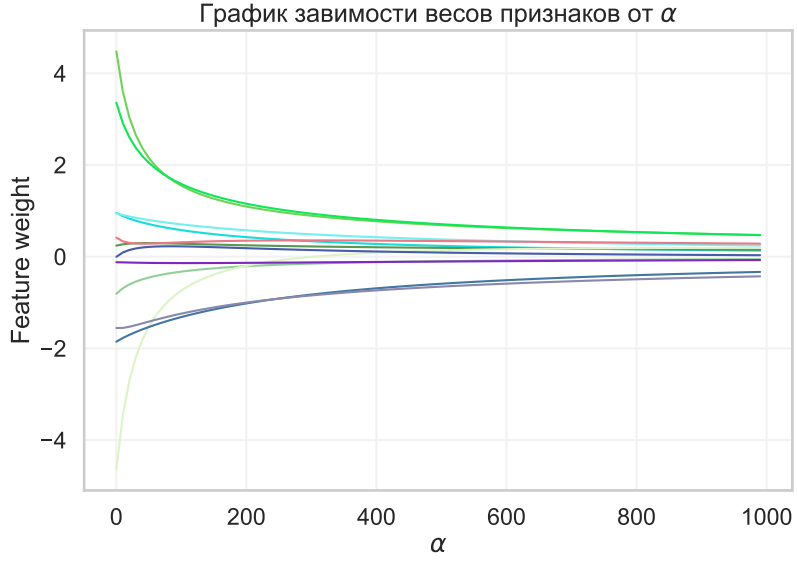
$$S_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f_j(\boldsymbol{x}_i) - \bar{f}_j)^2} \text{ — среднеквадратичное отклонение.}$$

Изменение параметра α при стандартизации

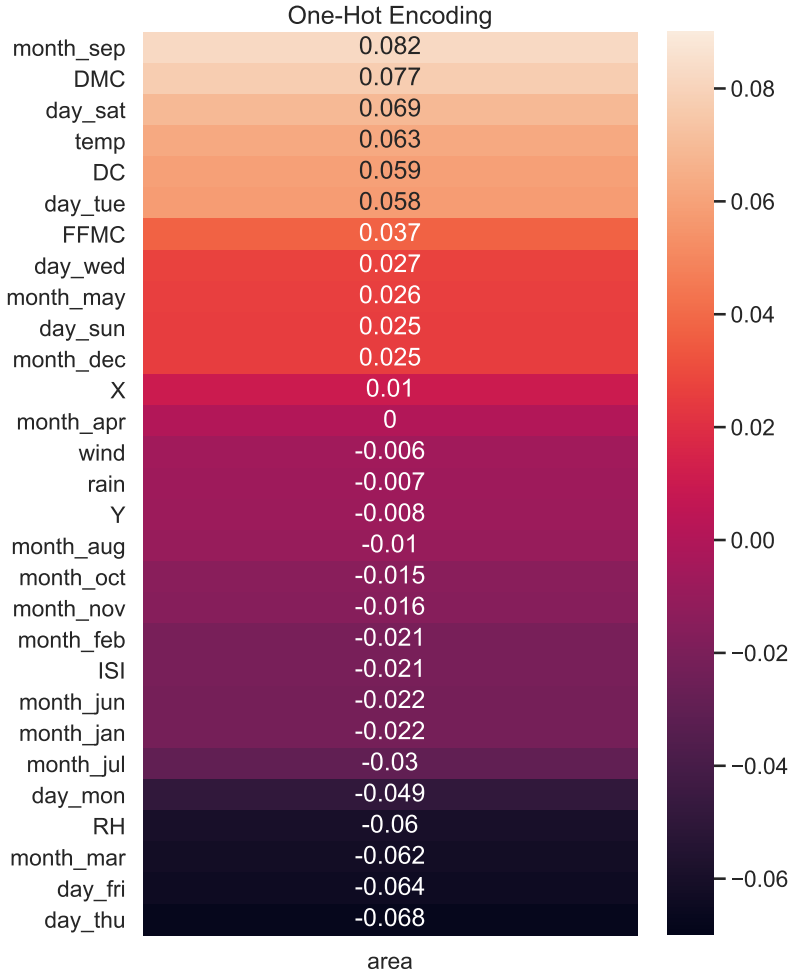


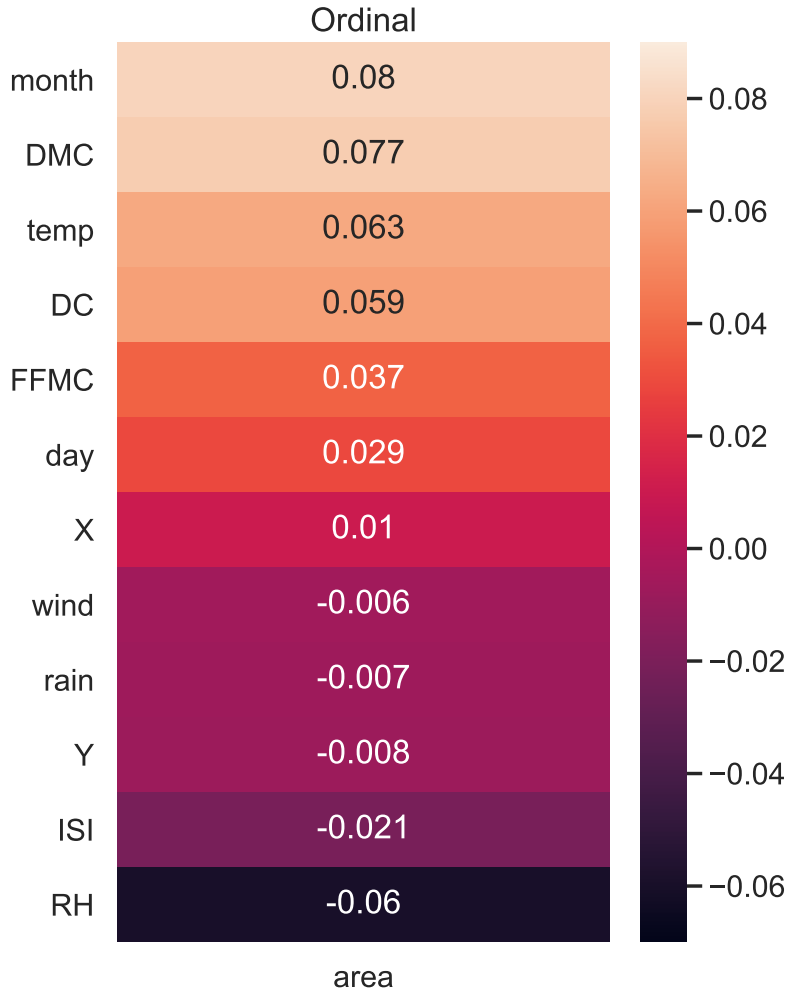






Отбор признаков

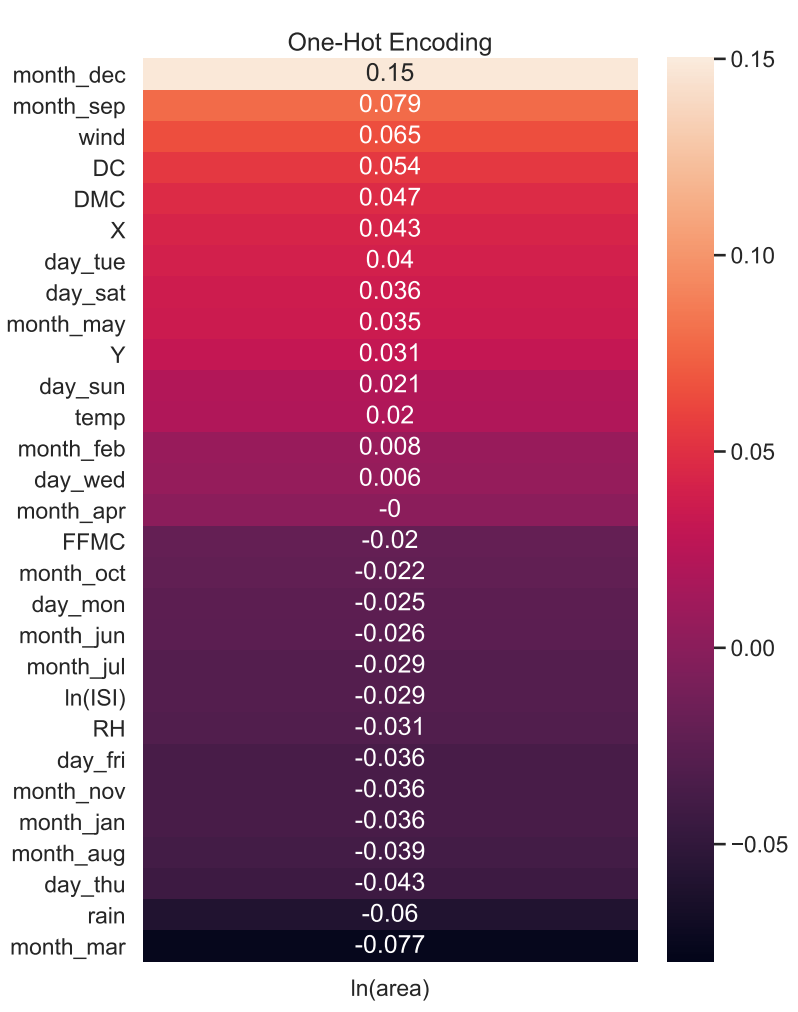


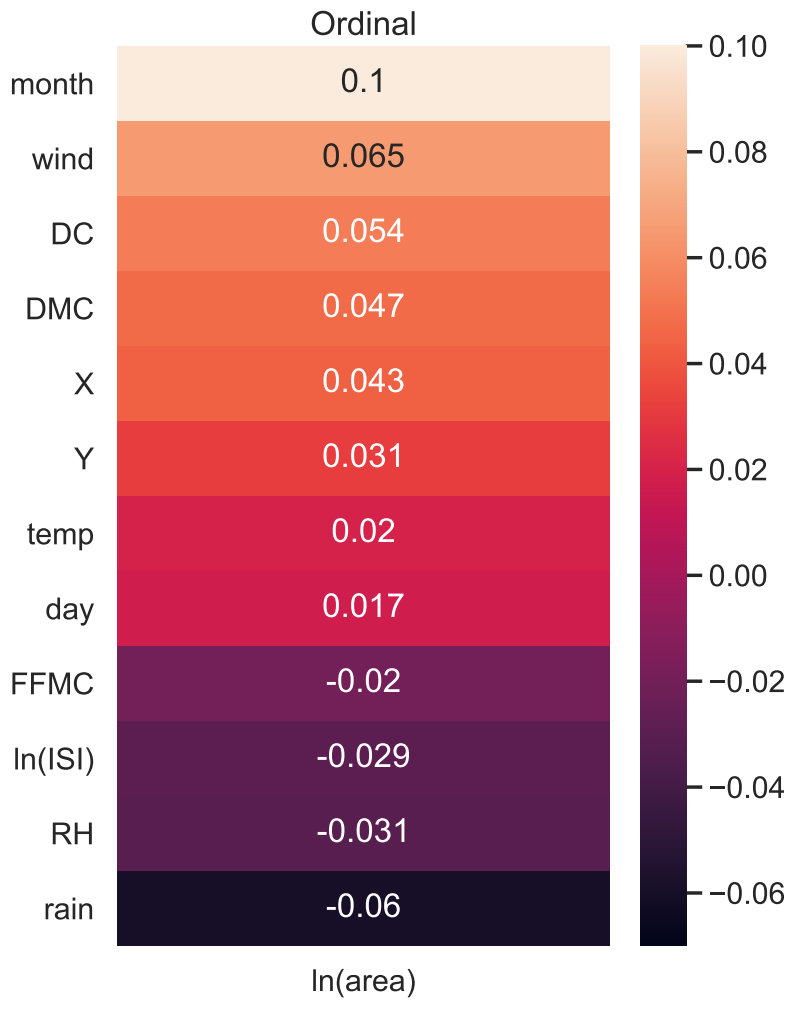


Преобразования признаков и ответов

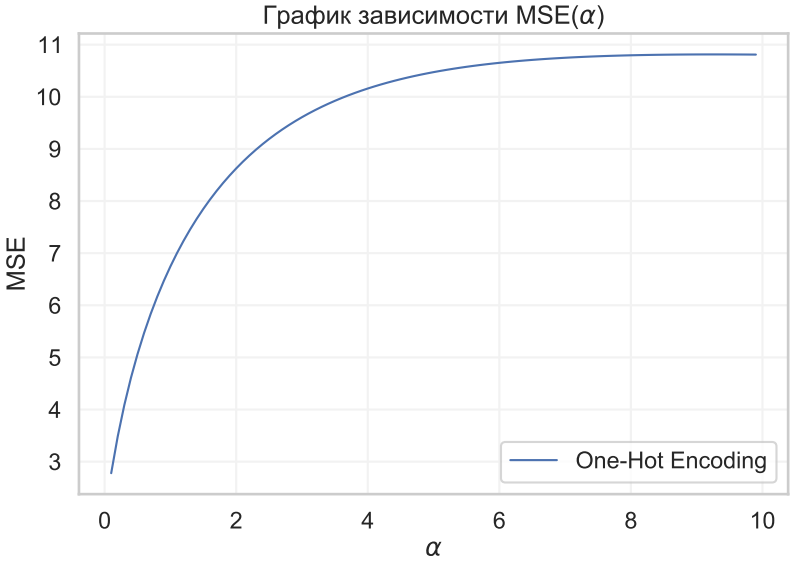
1. rain — номинальный
2. FFMC ≥ 75
3. ISI $\rightarrow \ln(ISI)$
4. area $\rightarrow \ln(1 + area)$

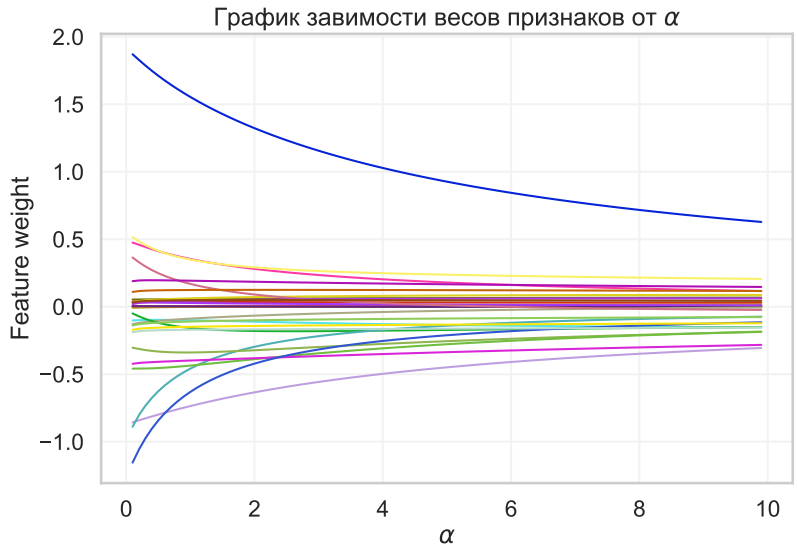
Взаимосвязь новых признаков и ответов

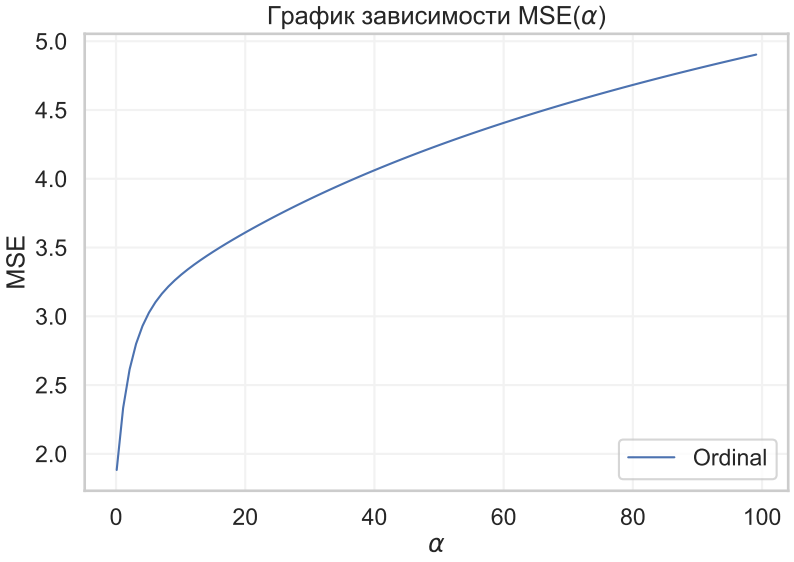


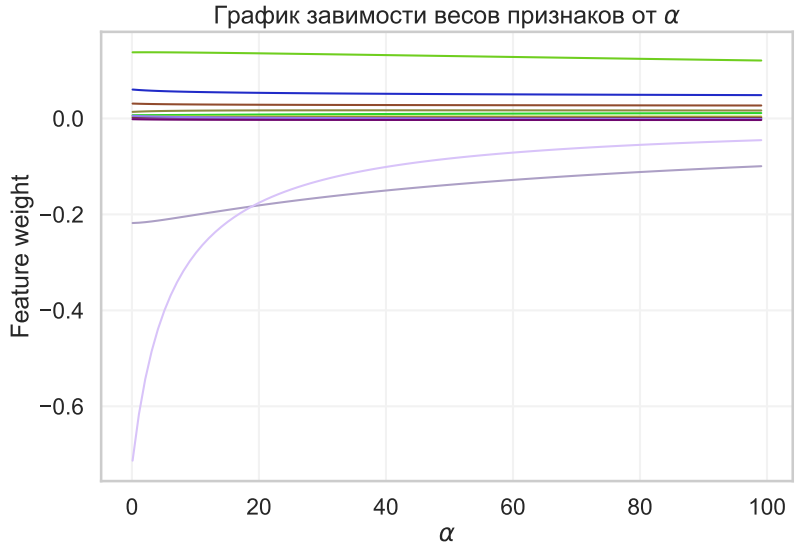


Изменение параметра α

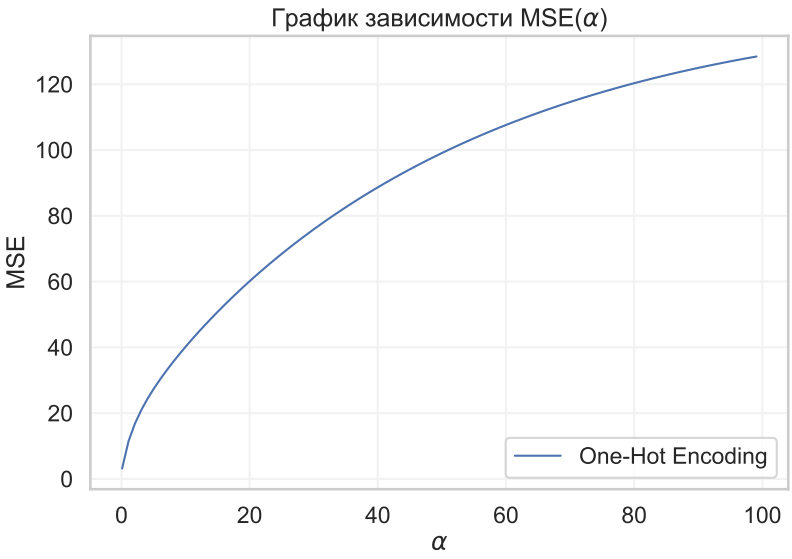


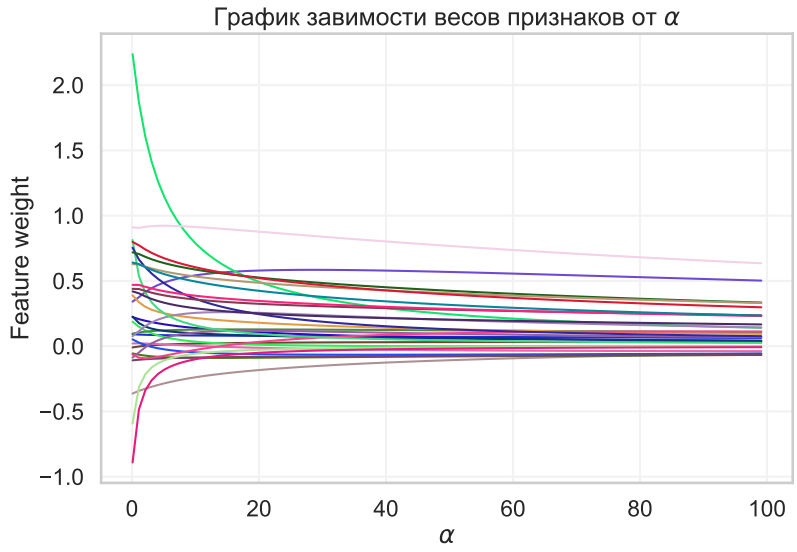


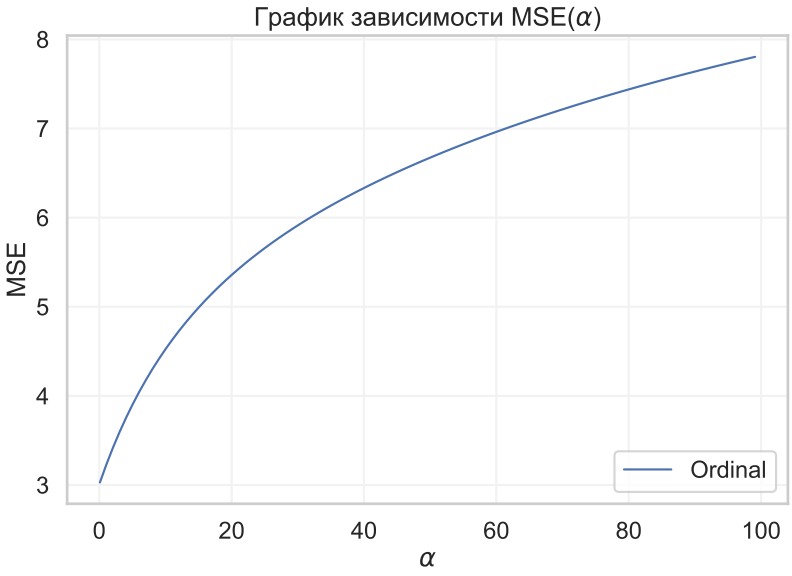


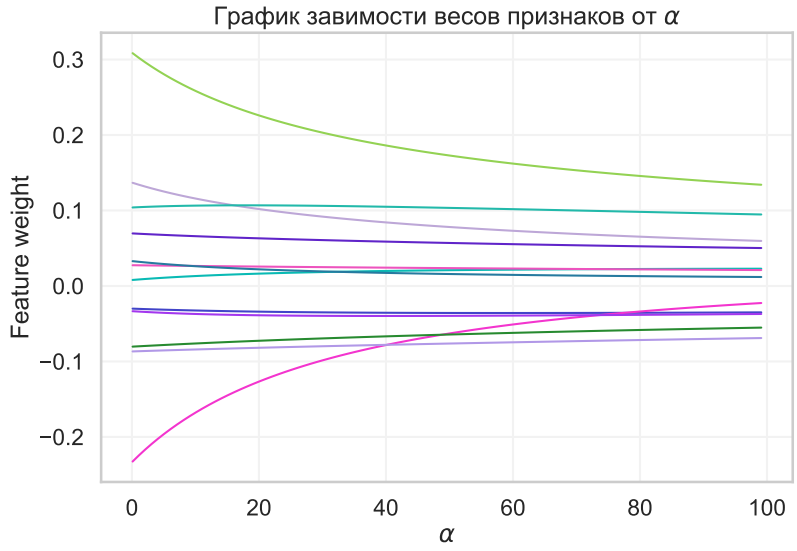


Изменение параметра α при стандартизации









Сравнение результатов

Таблица 1: Лучшее значение MSE на кросс-валидации

Стандартизация \ Преобразование	До	После
−	616,50	1,82
+	624,78	1,89

Стоит отметить, что после преобразования ответами являются $\ln(1 + area)$.

Наиболее значимые признаки

Значимость признаков при решении задачи лучше всего оценивается на данных после преобразования. Таковыми являются:

- month — месяц года
- wind — скорость ветра
- rain — количество осадков
- DC и DMC — индексы засухи и влажности почвы