**Лабораторная работа № 3 – Регрессионный анализ данных**

Цель лабораторной работы – получение навыков работы с методами регрессии

1. Изучить набор данных. Создать описание набора данных на русском языке. Описать признаки, используемые в наборе данных (включить полученные описания в начало ноутбука).
2. Проанализируйте колонки в наборе данных и сделайте вывод в какой колонке находится целевое значение, которое должна прогнозировать модель. Анализ отразите в ноутбуке.
3. Удалите дубликаты строк в наборе данных; приведите размер набора данных до и после данной операции;
4. Удалите выбросы из числовых данных. Используйте критерий 1.5-межквантильного интервала. Приведите диаграммы BoxPlot до и после удаления выбросов.
5. Оцените корреляцию между числовыми признаками, постройте матрицу корреляции. Найдите признаки, которые коррелируют с наибольшим числом других и выпишите их >|0.5| (все кроме целевой колонки).
6. Проанализируйте колонки на наличие нулевых значений. Выведите количество нулевых по каждой колонке.
7. Проанализируйте почему в той или иной колонке могут присутствовать нулевые значения.
8. Выполните удаление столбцов с большим количеством нулевых значений, где это возможно, используйте Imputers для числовых колонок.
9. Выполните масштабирование количественных (числовых) признаков:
   1. Постройте диаграммы BoxPlot для признаков до масштабирования.
   2. Выберите способ масштабирования (например, нормализацию или стандартизацию MinMaxScaler, StandardScaler)
   3. Выполните масштабирование выбранным методом
   4. Постройте диаграммы BoxPlot для признаков после масштабирования.
10. Выполните замену категориальных признаков; выберите и **обоснуйте способ замены** (one hot encoder, label encoder);
11. Решите поставленную задачу регрессии см. пункт 2.
    1. Разделите набор данных на две выборки: обучение и тест
    2. По каждому методу регрессии из задания изучите возможные для изменения параметры и их значения
    3. При подборе параметров модели регрессии используйте перекрёстную проверку (GridSearch CrossValidate). Обоснуйте выбор метрики для параметра refit в GridSearchCV.
    4. Вычислите точность решения задачи, вычислив разницу между реальным значением и предсказанным на тестовой выборке (MAE или MSE).
    5. Вычислите коэффициент корреляции (Пирсона, Спирмена) между реальным значением и предсказанным. Проанализируйте значение p-value.

Кодировка моделей:

1 – линейная регрессия без использования регуляризаторов;

2 – линейная регрессия, LASSO регуляризатор;

3 – линейная регрессия, RIDGE регуляризатор;

Варианты заданий:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Вариант | Набор данных / Модели | Вариант | Набор данных / Модели | Вариант | Набор данных / Модели |
| 1 | 1 / 1, 3 | 2 | 2 / 1, 3 | 3 | 3 / 1, 3 |
| 4 | 1 / 1, 2 | 5 | 2 / 1, 2 | 6 | 3 / 1, 2 |
| 7 | 1 / 2, 3 | 8 | 2 / 2, 3 | 9 | 3 / 2, 3 |
| 10 | 1 / 1, 2 | 11 | 2 / 1, 2 | 12 | 3 / 1, 2 |
| 13 | 1 / 2, 3 | 14 | 2 / 2, 3 | 15 | 3 / 2, 3 |
| 16 | 4 / 1, 3 | 17 | 5 / 1, 3 | 18 | 6 / 1, 3 |
| 19 | 4 / 1, 2 | 20 | 5 / 1, 2 | 21 | 6 / 1, 2 |
| 22 | 4 / 2, 3 | 23 | 5 / 2, 3 | 24 | 6 / 2, 3 |
| 25 | 4 / 1, 3 | 26 | 5 / 1, 3 | 27 | 6 / 1, 3 |
| 28 | 4 / 1, 2 | 29 | 5 / 1, 2 | 30 | 6 / 1, 2 |
| 31 | 7 / 1, 3 | 32 | 8 / 1, 3 |  |  |