

Домашнее задание по теме «Линейная регрессия»

Формулировка задания

Выполните подготовку данных для решения задачи регрессии. Проведите регрессию методом наименьших квадратов `statsmodel` и линейной регрессии `scikit-learn`. Качество оценить минимум по 3 критериям качества для регрессии: MAE, MSE, RMSE, MAPE и др.

Для регрессии и оценки качества использовать библиотеки `scikit-learn` и `statsmodel`.

Результирующий код должен быть читаемым, с единой системой отступов и адекватными названиями переменных.

Описание плана работы

- 1) Загрузите данные из дополнительных материалов или по ссылке: <https://www.kaggle.com/competitions/playground-series-s3e11/data>
- 2) Проведите EDA(Exploratory Data Analysis), то есть для каждой переменной посчитайте:
 - a) Долю пропусков
 - b) Максимальное и минимальное значение
 - c) Среднее значение
 - d) Медиану
 - e) Дисперсию
 - f) Квантиль 0.1 и 0.9
 - g) Квартиль 1 и 3
- 3) Подготовка датасета к построению моделей ML
 - a. Столбец “cost” является целевой переменной. Посмотреть как данные распределены. Если требуется, то обработать выбросы.
 - b. Провести анализ и обработку пропусков (либо заменить, либо удалить)
 - c. Провести анализ и обработку выбросов (либо заменить, либо удалить)
 - d. Рассчитать матрицу парных корреляций. Принять решение какие столбцы выбрать входными (независимые переменные) для модели линейной регрессии.
 - e. Разделить набор данных на обучающую и тестовую выборки

- 4) Обучить 2 модели. Из библиотеки statsmodels выбрать модель OLS. Из библиотеки scikit-learn выбрать модель LinearRegression.
- 5) Оценить качество алгоритмов, выбрать самый оптимальный алгоритм.
- 6) Из модели LinearRegression выгрузить значения коэффициентов модели. Визуализировать веса признаков и сделать вывод по модели.

Перечень необходимых инструментов

- Python
- scikit-learn
- statsmodels
- pandas
- matplotlib
- seaborn
- venv
- Jupiter Notebook
- IDE VS Code
- GigaIDE

Форма предоставления результата

1. В поле ссылки загрузить ссылку на удаленный репозиторий с доступом для наставника.
2. В поле файла загрузить архив с папкой, в которой разместить отчет со скриншотами по заданию и решение задачи. Решение должно быть представлено в формате .ipynb или .py.

Шкала оценивания

- 1.0 – отлично
- 0.7–0.9 – хорошо
- 0.5–0.6 – удовлетворительно
- Менее 0.5 – задание не выполнено