Лабораторная работа 3

Описание датасета

Датасет состоит из двух файлов: `TrainData3.csv` и `TestData3.csv`, содержащих тренировочные и тестовые данные соответственно. Каждый файл включает в себя анонимизированные характеристики клиентов некоторой компании и целевую переменную, указывающую, покинет ли клиент компанию.

- TrainData3.csv: Тренировочный набор данных, содержащий 1000 записей.

- TestData3.csv: Тестовый набор данных, содержащий 300 записей.

Каждый файл включает 15 столбцов, из которых 14 — это анонимизированные признаки клиентов, а 15-й столбец — целевая переменная.

1. feature_0 до feature_13: Анонимизированные числовые признаки клиентов. Значения этих признаков варьируются от 0 до 100.

2. target: Целевая переменная (0 или 1). Показывает, покинет ли клиент компанию:

  - `1`: Клиент покинет компанию

  - `0`: Клиент останется в компании

Задача оттока клиентов

Часть 1: Работа с пропусками

Задание 1

Проверьте, есть ли в тренировочных и тестовых данных пропуски? Укажите количество столбцов тренировочной выборки, имеющих пропуски.

Задание 2

a) В столбце с наибольшим количеством пропусков заполните пропуски средним значением по столбцу. В ответ запишите значение вычисленного среднего. Ответ округлите до десятых.

b) Найдите строки в тренировочных данных, где пропуски стоят в столбце с наименьшим количеством пропусков. Удалите эти строки. Сколько строк вы удалили?

Часть 2: Предобработка данных

Задание 3

Выполните следующие пункты только по таблице train.

a) Сколько столбцов в таблице (не считая target) содержат меньше 5 различных значений?

b) Вычислите долю ушедших из компании клиентов, для которых значение признака 2 больше среднего значения по столбцу, а значение признака 13 меньше медианы по столбцу. Ответ округлите до сотых.

Часть 3: Обучение модели

Задание 4

a) Разбейте тренировочные данные на целевой вектор y, содержащий значения из столбца target, и матрицу объект-признак X, содержащую

остальные признаки. Обучите на этих данных логистическую регрессию из sklearn (LogisticRegression) с параметрами по умолчанию. Выведите среднее значение метрики f1-score алгоритма на кросс-валидации с тремя фолдами. Ответ округлите до сотых.

При объявлении модели фиксируйте random_state = 42.

Комментарий: параметры по умолчанию можете оставить дефолтными

Задание 5

a) Подберите значение константы регуляризации C в логистической регрессии, перебирая гиперпараметр от 0.001 до 100 включительно, проходя по степеням 10. Для выбора C примените перебор по сетке по тренировочной выборке (GridSearchCV из библиотеки sklearn.model_selection) с тремя фолдами и метрикой качества - f1-score. Остальные параметры оставьте по умолчанию. В ответ запишите наилучшее среди искомых значение C.

При объявлении модели фиксируйте random_state = 42.

Комментарий: параметры по умолчанию можете оставить дефолтными

b) Добавьте в тренировочные и тестовые данные новый признак 'NEW', равный произведению признаков '7' и '11'. На тренировочных данных с новым признаком заново с помощью GridSearchCV (с тремя фолдами и метрикой качества - f1-score) подберите оптимальное значение C (перебирайте те же значения C, что и в предыдущих заданиях), в ответ напишите наилучшее качество алгоритма (по метрике f1-score), ответ округлите до сотых.

При объявлении модели фиксируйте random_state = 42.

c) Теперь вы можете использовать любую модель машинного обучения для решения задачи. Также можете делать любую другую обработку признаков. Ваша задача - получить наилучшее качество по метрике F1-Score на тестовых данных.

Лучший результат будет можно будет добавить в викторину в Телеграмм, победитель получит +2 балла к оценке