

Машинное обучение: Сердечные импутации

Предыстория

Компания Lemon AI получила контракт с известным кардиологическим центром в Риге. Ивану поручили упорядочить исторические данные пациентов больницы, чтобы обучить модель, предсказывающую наличие сердечно-сосудистых заболеваний.

Однако сервер базы данных больницы был очень старым и работал со сбоями. Во время экспорта данных около 10% значений были повреждены и потеряны (заменены на `NaN`).

Более того, главный кардиолог была очень строга: она считала, что не выявить пациента с болезнью сердца в **3 раза хуже**, чем ошибочно поставить диагноз здоровому человеку.

«Если мы упустим одного больного пациента, это риск для жизни», — объяснила она Ивану. «Ложноположительный результат — это просто дополнительные анализы. Я считаю, ложноотрицательный результат в 3 раза хуже, чем ложноположительный».

Ивану должен был очистить поврежденные данные и построить модель, удовлетворяющую строгим требованиям безопасности главного кардиолога. Это была его самая ответственная задача — с реальными последствиями.

Задача

Вам дан набор показателей здоровья пациентов:

- `x_train.csv`: Признаки (features) для обучающего набора. Содержит пропущенные значения.
- `y_train.csv`: Целевые метки (labels) для обучающего набора (`1` = есть болезнь сердца, `0` = нет болезни).
- `x_test.csv`: Признаки (features) для тестового набора. Содержит пропущенные значения.

Ваша задача — предсказать целевые значения для тестового набора.

Подсказки

- Предобработка: Изучите модуль `sklearn.impute` для обработки значений `NaN` в CSV-файлах.
- Оптимизация: Внимательно изучите формулу оценивания.
- Вам может пригодиться это руководство от scikit-learn: <https://scikit-learn.org/stable/modules/impute.html>

Оценивание

Решения оцениваются на основе метрики взвешенной точности (weighted accuracy) по формуле:

$$Score = \frac{3 \cdot TP + TN}{3 \cdot (TP + FN) + (TN + FP)}$$

Где:

- **TP (True Positive — Истинно положительный ответ):** Вы предсказали болезнь, и она действительно есть. (Вес x3)
- **TN (True Negative — Истинно отрицательный ответ):** Вы предсказали, что пациент здоров, и он действительно здоров.
- **FP (False Positive — Ложноположительный ответ):** Вы предсказали болезнь, но пациент здоров.
- **FN (False Negative — Ложноотрицательный ответ):** Вы предсказали, что пациент здоров, но у него на самом деле болезнь. (Это снижает балл сильнее всего, так как уменьшает наиболее значимое слагаемое в числителе).

По сути, балл равен точности (ассигасу) на наборе данных, где каждый положительный пациент продублирован 3 раза.

Итоговый балл за задачу рассчитывается как: $(\text{score} - \text{baseline}) / (\text{max_score} - \text{baseline})$.