

УДК 004.04Python

ББК 32.372

Г90

Груздев А. В.

Г90 Предварительная подготовка данных в Python. Том 2: План, примеры и метрики качества. – М.: ДМК Пресс, 2023. – 814 с.: ил.

ISBN 978-5-93700-177-1

В двухтомнике представлены материалы по применению классических методов машинного обучения в различных промышленных задачах. Во втором томе рассматривается сам процесс предварительной подготовки данных, а также некоторые метрики качества и ряд полезных библиотек и фреймворков (H2O, Dask, Docker, Google Colab).

Издание рассчитано на специалистов по анализу данных, а также может быть полезно широкому кругу специалистов, интересующихся машинным обучением.

УДК 004.04Python

ББК 32.372

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но, поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

Оглавление

Введение	7
ЧАСТЬ 4. ПЛАН ПРЕДВАРИТЕЛЬНОЙ ПОДГОТОВКИ ДАННЫХ	8
1. Введение	8
2. Формирование выборки	10
2.1. Генеральная и выборочная совокупности	10
2.2. Характеристики выборки.....	10
2.3. Детерминированные и вероятностные выборки	12
2.4. Виды, методы и способы вероятностного отбора	13
2.5. Подходы к определению необходимого объема выборки	14
3. Определение «окна выборки» и «окна созревания»	28
4. Определение зависимой переменной	32
5. Загрузка данных из CSV-файлов и баз данных SQL.....	33
6. Удаление бесполезных переменных, переменных «из будущего», переменных с юридическим риском.....	39
7. Преобразование типов переменных и знакомство со шкалами переменных	41
7.1. Количественные (непрерывные) шкалы.....	41
7.2. Качественные (дискретные) шкалы.....	43
8. Нормализация строковых значений	45
9. Обработка дублирующихся наблюдений.....	61
10. Обработка редких категорий	62
11. Появление новых категорий в новых данных	69
12. Импутация пропусков.....	70
12.1. Способы импутации количественных и бинарных переменных.....	70

12.2. Способы импутации категориальных переменных	71
12.3. Практика	73
13. Обработка выбросов.....	90
14. Описательные статистики	94
14.1. Пифагорейские средние, медиана и мода	94
14.2. Квантиль	95
14.3. Дисперсия и стандартное отклонение	96
14.4. Корреляция и ковариация	97
14.5. Получение сводки описательных статистик в библиотеке pandas	102
15. Нормальное распределение.....	104
15.1. Знакомство с нормальным распределением	104
15.2. Коэффициент островершинности, коэффициент эксцесса и коэффициент асимметрии	107
15.3. Гистограмма распределения и график квантиль–квантиль.....	111
15.4. Вычисление коэффициента асимметрии и коэффициента эксцесса, построение гистограммы и графика квантиль–квантиль для подбора преобразований, максимизирующих нормальность	112
15.5. Подбор преобразований, максимизирующих нормальность для правосторонней асимметрии	116
15.6. Подбор преобразований, максимизирующих нормальность для левосторонней асимметрии.....	128
15.7. Преобразование Бокса–Кокса	129
16. Конструирование признаков	135
16.1. Статическое конструирование признаков исходя из предметной области	135
16.2. Статическое конструирование признаков исходя из алгоритма	170
16.3. Динамическое конструирование признаков исходя из особенностей алгоритма	290
16.4. Конструирование признаков для временных рядов	297
17. Отбор признаков	433
17.1. Методы-фильтры	436
17.2. Применение метода-фильтра и встроенного метода для отбора признаков (на примере соревнования BNP Paribas Cardif Claims Management с Kaggle)	444
17.3. Комбинирование нескольких методов для отбора признаков (на примере соревнования Porto Seguro’s Safe Driver Prediction с Kaggle)	451
18. Стандартизация.....	475
19. Собираем все вместе	486

ЧАСТЬ 5. МЕТРИКИ ДЛЯ ОЦЕНКИ КАЧЕСТВА МОДЕЛИ....514

1. Бинарная классификация.....514

1.1. Отрицательный и положительный классы, порог отсечения	514
1.2. Матрица ошибок	514
1.3. Доля правильных ответов, правильность (accuracy)	517
1.4. Чувствительность (sensitivity).....	519
1.5. Специфичность (specificity)	521
1.6. 1 – специфичность (1 – specificity)	522
1.7. Сбалансированная правильность.....	523
1.8. Точность (Precision).....	524
1.9. Сравнение точности и чувствительности (полноты)	525
1.10. F-мера (F-score, или F-measure)	526
1.11. Варьирование порога отсечения.....	532
1.12. Коэффициент Мэттьюса (Matthews correlation coefficient или MCC).....	536
1.13. Каппа Коэна (Cohen's kappa).....	540
1.14. ROC-кривая (ROC curve) и площадь под ROC-кривой (AUC-ROC).....	542
1.15. PR-кривая (PR curve) и площадь под PR-кривой (AUC-PR)	603
1.16. Кривая Лоренца (Lorenz curve) и коэффициент Джини (Gini coefficient).....	616
1.17. CAP-кривая (CAP curve).....	620
1.18. Статистика Колмогорова–Смирнова (Kolmogorov–Smirnov statistic)	623
1.19. Биномиальный тест (binomial test)	626
1.20. Логистическая функция потерь (logistic loss)	628

2. Регрессия.....634

2.1. R^2 , коэффициент детерминации (R-square, coefficient of determination)	634
2.2. Метрики качества, которые зависят от масштаба данных (RMSE, MSE, MAE, MdAE, RMSLE, MSLE)	643
2.3. Метрики качества на основе процентных ошибок (MAPE, MdAPE, sMAPE, sMdAPE, WAPE, WMAPE, RMSPE, RMdSPE).....	656
2.4. Метрики качества на основе относительных ошибок (MRAE, MdRAE, GMRAE)	689
2.5. Относительные метрики качества (RelMAE, RelRMSE)	697
2.6. Масштабированные ошибки (MASE, MdASE).....	698
2.7. Критерий Диболда–Мариано	705

ЧАСТЬ 6. ДРУГИЕ ПОЛЕЗНЫЕ БИБЛИОТЕКИ И ПЛАТФОРМЫ707

1. Библиотеки байесовской оптимизации hyperopt, scikit-optimize и optuna707

1.1. Недостатки обычного поиска по сетке и случайного поиска по сетке.....	707
1.2. Знакомство с байесовской оптимизацией.....	708
1.3. Последовательная оптимизация по модели (Sequential model-based optimization – SMBO)	710
1.4. Hyperopt.....	716
1.5. Scikit-Optimize	727
1.6. Optuna	732
2. Docker	742
2.1. Введение	742
2.2. Запуск контейнера Docker.....	743
2.3. Создание контейнера Docker с помощью Dockerfile	744
3. Библиотека H2O	749
3.1. Установка пакета h2o для Python.....	749
3.2. Запуск кластера H2O	749
3.3. Преобразование данных во фреймы H2O	750
3.4. Знакомство с содержимым фрейма.....	751
3.5. Определение имени зависимой переменной и списка имен признаков	753
3.6. Построение модели машинного обучения.....	753
3.7. Вывод модели	754
3.8. Получение прогнозов	758
3.9. Построение ROC-кривой и вычисление AUC-ROC.....	759
3.10. Поиск оптимальных значений гиперпараметров по сетке	760
3.11. Извлечение наилучшей модели по итогам поиска по сетке.....	762
3.12. Класс H2OAutoML.....	762
3.13. Применение класса H2OAutoML в библиотеке scikit-learn	771
4. Библиотека Dask	783
4.1. Общее знакомство	783
4.2. Машинное обучение с помощью библиотеки dask-ml.....	792
4.3. Построение конвейера в Dask	800
5. Google Colab.....	804
5.1. Общее знакомство	804
5.2. Регистрация и создание папки проекта	804
5.3. Подготовка блокнота Colab	809