

УДК 519.7:004.9IBM SPSS Statistics
ББК 21.18с
Г90

Груздев А. В.

Г90 Прогнозное моделирование в IBM SPSS Statistics, R и Python: метод деревьев решений и случайный лес. – М.: ДМК Пресс, 2018. – 642 с.: ил.

ISBN 978-5-97060-539-4

Данная книга представляет собой практическое руководство по применению метода деревьев решений и случайного леса для задач сегментации, классификации и прогнозирования. Каждый раздел книги сопровождается практическим примером. Кроме того, книга содержит программный код SPSS Syntax, R и Python, позволяющий полностью автоматизировать процесс построения прогнозных моделей. Автор обобщены лучшие практики использования деревьев решений и случайного леса от таких компаний, как Citibank N.A., Transunion и DBS Bank.

Издание будет интересно маркетологам, риск-аналитикам и другим специалистам, занимающимся разработкой и внедрением прогнозных моделей.

УДК 519.7:004.9IBM SPSS Statistics
ББК 21.18с

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.

ISBN 978-5-97060-539-4

© Груздев А. В., 2018
 © Оформление, издание, ДМК Пресс, 2018

Содержание

От рецензента	10
Предисловие	11
Глава 1. Введение в метод деревьев решений	14
1.1. Введение в методологию деревьев решений	14
1.2. Преимущества и недостатки деревьев решений	19
1.3. Задачи, выполняемые с помощью деревьев решений	20
Вопросы к главе 1	22
Часть I. ПОСТРОЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ И СЛУЧАЙНОГО ЛЕСА В IBM SPSS STATISTICS	23
Глава 2. Основы прогнозного моделирования с помощью деревьев решений CHAID	24
2.1. Запуск процедуры Деревья классификации	24
2.2. Четыре метода деревьев решений	26
2.3. Шкалы переменных	29
2.4. Определение необходимого размера выборки	31
2.5. Знакомство с методом CHAID	32
2.5.1. Описание алгоритма	32
2.5.2. Немного о тесте хи-квадрат	35
2.5.3. Немного об F-тесте	37
2.5.4. Способы объединения категорий предикторов	38
2.5.5. Поправка Бонферрони	38
2.5.6. Иллюстрация работы CHAID на конкретном примере	38
2.6. Построение и интерпретация дерева классификации CHAID	43
2.6.1. Сводка для модели	45
2.6.2. Диаграмма дерева	46
2.6.3. Выигрыши для узлов	48
2.6.4. Таблицы классификации и риска	49
2.7. Работа с прогнозами модели	51
2.7.1. Получение результатов классификации	51
2.7.2. Сохранение прогнозов модели в файле данных	52
2.7.3. Самостоятельное построение таблицы классификации и изменение порогового значения вероятности	57
2.8. Анализ ROC-кривой	66
2.8.1. Терминология анализа ROC-кривой	66
2.8.2. Оценка дискриминирующей способности модели и выбор порогового значения с помощью ROC-кривой	73
2.9. Диагностика качества модели	79
2.9.1. Обобщающая способность, переобучение и недообучение	79
2.9.2. Методы проверки модели	80
2.9.3. Общие правила интерпретации результатов проверки	83
2.9.4. Методы проверки модели, реализованные в процедуре Деревья классификации	85

2.9.5. Практическое применение методов проверки в процедуре Деревья классификации	86
2.9.6. Самостоятельное разбиение набора данных на обучающую и контрольную выборки для осуществления проверки	97
2.10. Дополнительные настройки вывода результатов	101
2.10.1. Настройки вывода дерева	101
2.10.2. Построение таблицы дерева	102
2.10.3. Настройки вывода статистик	103
2.10.4. Построение таблиц выигрышей для узлов и процентилей.....	105
2.10.5. Настройки вывода графиков.....	107
2.10.6. Построение графиков выигрышей, индексов и откликов.....	109
2.10.7. Настройки вывода правил классификации.....	111
2.10.8. Применение правил классификации к новому набору данных	112
2.11. Построение дерева регрессии CHAID.....	122
2.12. Использование принудительной переменной расщепления	127
Выводы и рекомендации	129
Вопросы к главе 2.....	130

Глава 3. Продвинутое моделирование

с помощью деревьев решений CHAID.....	133
3.1. Построение деревьев CHAID с измененными критериями.....	133
3.1.1. Настройка правил остановки	133
3.1.2. Построение деревьев CHAID с измененными правилами остановки	134
3.1.3. Настройка статистических тестов для разбиения узлов и объединения категорий предикторов.....	140
3.1.4. Построение дерева CHAID с измененными статистическими тестами.....	141
3.1.5. Настройка обработки количественных предикторов	142
3.1.6. Построение дерева CHAID с измененным числом интервалов для количественных предикторов	143
3.2. Метод Исчерпывающий CHAID	144
3.3. Обзор параметров деревьев решений.....	145
3.4. Работа с пропусками в методе CHAID	147
3.4.1. Настройка обработки пропущенных значений	147
3.4.2. Построение дерева CHAID на основе данных, содержащих пропуски	150
3.5. Работа со стоимостями ошибочной классификации в методе CHAID.....	151
3.5.1. Настройка стоимостей ошибочной классификации	151
3.5.2. Построение дерева CHAID с измененными стоимостями ошибочной классификации	154
3.6. Работа с прибылями в методе CHAID.....	157
3.6.1. Настройка прибылей	157
3.6.2. Построение дерева CHAID с заданными значениями прибыли.....	158
3.7. Работа со значениями	162
3.8. Применение метода CHAID для биннинга переменных (на примере конкурсной задачи ОТП Банка).....	165
3.8.1. Преимущества и недостатки биннинга	165
3.8.2. Предварительная подготовка данных	167
3.8.3. Определение важности переменных с помощью случайного леса.....	184
3.8.4. Анализ мультиколлинеарности	187
3.8.5. Выполнение биннинга переменных на основе CHAID.....	188

3.8.6. Построение моделей логистической регрессии на основе исходных предикторов и предикторов, категоризированных с помощью CHAID	194
3.8.7. Выполнение биннинга переменных с помощью процедуры Оптимальная категоризация	199
3.8.8. Построение модели логистической регрессии на основе оптимально категоризированных предикторов.....	202
3.8.9. Преобразование количественных переменных для максимизации нормальности	203
3.8.10. Построение модели логистической регрессии с использованием CHAID и преобразования корня третьей степени.....	207
3.9. Построение ансамбля логистической регрессии и дерева CHAID (на примере конкурсной задачи Tinkoff Data Science Challenge).....	208
Выводы и рекомендации	218
Вопросы к главе 3.....	219

Глава 4. Построение деревьев решений CRT и QUEST

4.1. Знакомство с методом CRT	220
4.1.1. Описание алгоритма	221
4.1.2. Мера Джини	222
4.1.3. Внутриузловая дисперсия	223
4.1.4. Метод отсеечения ветвей на основе меры стоимости-сложности	224
4.1.5. Обработка пропущенных значений.....	225
4.1.6. Иллюстрация работы CRT на конкретном примере.....	225
4.2. Построение дерева классификации CRT.....	228
4.3. Построение дерева CRT с измененными критериями	231
4.3.1. Настройка мер неоднородности для отбора предикторов и расщепления узлов	232
4.3.2. Настройка отсеечения ветвей.....	233
4.3.3. Построение дерева CRT с последующим отсечением ветвей	234
4.3.4. Настройка суррогатов для обработки пропущенных значений	235
4.3.5. Построение дерева CRT на основе данных, содержащих пропуски	236
4.4. Вывод важности предикторов.....	239
4.5. Работа с априорными вероятностями в методе CRT.....	240
4.5.1. Настройка априорных вероятностей	240
4.5.2. Построение дерева CRT с измененными априорными вероятностями	241
4.6. Знакомство с методом QUEST.....	243
4.6.1. Описание алгоритма	244
4.6.2. Метод отсеечения ветвей на основе меры стоимости-сложности	246
4.7. Построение дерева классификации QUEST	246
4.8. Сравнение метода QUEST с другими методами деревьев решений	248
4.9. Построение дерева QUEST с измененными критериями.....	249
4.9.1. Настройка статистических тестов для отбора предикторов.....	250
4.9.2. Построение дерева QUEST с последующим отсечением ветвей.....	250
Выводы и рекомендации	252
Вопросы к главе 4.....	252

Глава 5. Редактор дерева

5.1. Просмотр диаграммы дерева в Редакторе	254
5.2. Просмотр содержимого узла в Редакторе.....	255

5.3. Настройка внешнего вида диаграммы дерева в Редакторе.....	256
5.4. Изменение ориентации диаграммы дерева в Редакторе.....	257
5.5. Настройка содержимого узла в Редакторе	257
5.6. Отбор наблюдений в Редакторе	258
5.7. Иллюстрация работы в Редакторе дерева на конкретном примере	259

Глава 6. Построение случайного леса 263

6.1. Введение в методологию случайного леса.....	263
6.1.1. Описание метода.....	263
6.1.2. Оценка качества модели.....	267
6.1.3. Настройка параметров случайного леса	270
6.1.4. Важность предикторов.....	271
6.1.5. Графики частной зависимости	273
6.1.6. Матрица близостей	275
6.1.7. Обработка пропущенных значений.....	276
6.1.8. Обнаружение выбросов	276
6.1.9. Преимущества и недостатки случайного леса.....	277
6.1.10. История создания метода	278
6.2. Знакомство с процедурой Оценка RanFor	278
6.3. Построение ансамбля деревьев классификации	282
6.4. Интерпретация результатов, полученных с помощью ансамбля деревьев классификации	286
6.4.1. Сводка для модели	286
6.4.2. Важность переменных.....	288
6.4.3 Частота использования переменных	288
6.4.4 Матрица ошибок прогнозов.....	289
6.4.5. График частоты ошибок.....	290
6.4.6. График важности переменных.....	291
6.4.7. Графики частной зависимости	291
6.4.8 Работа с набором прогнозов.....	294
6.5. Проверка построенного ансамбля деревьев классификации на контрольной выборке и применение его к новым данным с помощью процедуры Прогноз RanFor.....	297
6.6. Построение ансамбля деревьев регрессии и интерпретация полученных результатов.....	303
6.7. Проверка построенного ансамбля деревьев регрессии на контрольной выборке и применение его к новым данным с помощью процедуры Прогноз RanFor	311
Выводы и рекомендации	315
Вопросы к главе 6.....	315

Часть II. ПОСТРОЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ И СЛУЧАЙНОГО ЛЕСА В R И PYTHON..... 318

Глава 7. Построение деревьев решений CHAID с помощью пакета R CHAID..... 319

7.1. Построение и интерпретация дерева классификации CHAID	319
7.1.1. Подготовка данных	319
7.1.2. Построение модели и работа с диаграммой дерева.....	321

7.1.3. Вычисление вероятностей классов и выбор оптимального порога.....	323
7.1.4. Получение спрогнозированных классов зависимой переменной.....	328
7.1.5. Сохранение прогнозов	329
7.1.6. Применение модели к новым данным	329
7.1.7. Проверка модели.....	330
7.2. Биннинг переменных	335
7.2.1. Биннинг в пакете rattle	335
7.2.2. Биннинг в пакете smbining.....	337
Выводы и рекомендации	344
Вопросы к главе 7.....	345

Глава 8. Построение деревьев решений CRT

с помощью пакета R rpart	346
8.1. Метод отсечения ветвей на основе стоимости-сложности с кросс-проверкой.....	346
8.2. Построение и интерпретация дерева классификации CRT	347
8.2.1. Подготовка данных	347
8.2.2. Построение модели и работа с диаграммой дерева.....	348
8.2.3. Прунинг дерева CRT	354
8.2.4. Вычисление вероятностей классов.....	356
8.2.5. Построение ROC-кривой и вычисление более точных оценок дискриминирующей способности.....	356
8.2.6. Сохранение спрогнозированных вероятностей	359
8.2.7. Применение модели к новым данным	359
8.3. Построение и интерпретация дерева регрессии CRT.....	361
8.3.1. Подготовка данных	361
8.3.2. Построение модели и работа с диаграммой дерева.....	362
Выводы и рекомендации	365
Вопросы к главе 8.....	365

Глава 9. Построение случайного леса с помощью пакета R randomForest

R randomForest	367
9.1. Построение ансамбля деревьев классификации	367
9.1.1. Подготовка данных	367
9.1.2. Построение модели и получение ООВ-оценки качества.....	369
9.1.3. Важности предикторов	374
9.1.4. Графики частной зависимости	375
9.1.5. Вычисление вероятностей классов.....	379
9.1.6. Оценка дискриминирующей способности модели с помощью ROC-кривой	380
9.1.7. Получение спрогнозированных классов зависимой переменной.....	383
9.1.8. График зазора прогнозов	385
9.2. Построение ансамбля деревьев регрессии.....	386
9.2.1. Подготовка данных	386
9.2.2. Построение модели и получение ООВ оценки качества.....	387
9.2.3. Важности предикторов	388
9.2.4. Графики частной зависимости	389
9.2.5. Работа с прогнозами и вычисление среднеквадратичной ошибки.....	391
9.2.6. Улучшение качества прогнозов.....	392
9.2.7. Вычисление коэффициента детерминации	393

9.2.8. Получение более развернутого вывода о качестве модели	394
9.3. Поиск оптимальных параметров случайного леса с помощью пакета caret	395
9.3.1. Схема оптимизации параметров, реализованная в пакете caret	395
9.3.2. Настройка условий оптимизации	396
9.3.3. Поиск оптимальных параметров для задачи регрессии	398
9.3.4. Поиск оптимальных параметров для задачи классификации	400
Выводы и рекомендации	410

Глава 10. Построение случайного леса с помощью пакета R ranger..... 411

10.1. Построение ансамбля деревьев классификации.....	411
10.2. Построение случайного леса вероятностей	433
10.3. Построение случайного леса выживаемости.....	442
Выводы и рекомендации	449

Глава 11. Построение распределенного случайного леса с помощью пакета R h2o..... 450

11.1. Решение задачи классификации.....	450
11.1.1. Подготовка данных	450
11.1.2. Построение модели и работа с результатами.....	455
11.1.3. Сохранение модели и применение к новым данным	466
11.1.4. Поиск оптимальных значений параметров с помощью решетчатого поиска.....	467
11.2. Решение задачи регрессии	478
Выводы и рекомендации	482

Глава 12. Построение случайного леса в Python..... 483

12.1. Знакомство с Python.....	483
12.1.1. Обзор основных инструментов Python, предназначенных для подготовки и анализа данных.....	483
12.1.2. Беспроblemная работа с программным кодом	490
12.2. Построение модели случайного леса и работа с полученными результатами.....	490
12.2.1. Подготовка данных в pandas.....	491
12.2.2. Параметры случайного леса и подгонка модели	500
12.2.3. Важности предикторов.....	505
12.2.4. Прогнозы модели и матрица ошибок	508
12.2.5. Отчет о результатах классификации: точность, полнота и F-мера	509
12.2.6. Построение ROC-кривой и выбор оптимального порога	511
11.2.7. Сравнение модели случайного леса с моделью дерева решений.....	514
12.3. Улучшение качества модели случайного леса.....	520
12.3.1. Методы перекрестной проверки, реализованные в scikit-learn.....	520
12.3.2. Поиск оптимальных параметров случайного леса.....	522
12.4. Построение распределенного случайного леса с помощью модуля H2O	541
12.4.1. Подготовка данных для построения стандартной модели случайного леса	541
12.4.2. Построение стандартной модели случайного леса	552
12.4.3. Применение стандартной модели случайного леса к новым данным.....	557
12.4.4. Подготовка данных для моделирования в H2O	560
12.4.5. Построение модели случайного леса с помощью класса H2ORandomForestEstimator.....	564

12.4.6. Сохранение модели случайного леса, построенной с помощью класса H2ORandomForestEstimator, и применение к новым данным	579
12.4.7. Улучшение качества моделей классов RandomForestClassifier и H2ORandomForestEstimator с помощью конструирования новых признаков	581
12.4.8. Выполнение решетчатого поиска с помощью класса H2OGridSearch	585
12.4.9. Улучшение качества модели с помощью стекинга	590
Выводы и рекомендации	598
Приложение 1. Предварительная подготовка данных в Python с помощью библиотеки pandas	599
Приложение 2. Предварительная подготовка данных в R	604
Приложение 3. Визуализация данных в Python с помощью библиотек matplotlib, seaborn и plotly	612
Приложение 4. Построение ROC-кривой и вычисление AUC вручную	616
Приложение 5. Декомпозиция прогнозов дерева решений и случайного леса с помощью питоновского пакета treeinterpreter для улучшения интерпретабельности	622
Ключи к вопросам	630
Библиографический список	631
Предметный указатель	633