

Настоящая книга подготовлена по личной инициативе автора.

Об авторе.

Фоменко Александр Александрович, кандидат технических наук, работает на фондовом рынке с момента его возникновения в России: чеки, ГКО, голубые фишки ММВБ, фьючерсы РТС, Форекс, преподавание инвестиций в ценные бумаги – полный набор. Последние несколько лет посвящены разработке эконометрических моделей, в результате чего и появилась данная книга.

Фоменко А.А.

Предсказываем тренды*.* С Rattle в мир моделей классификации. - 2014. 383 c.

Книга является практическим руководством по обучению моделей предсказаниям трендов на рынке Форекс. Берем значения исходных данных – котировок, индикаторов, макроэкономических данных, просто новостей, и разбиваем все это на несколько классов, в книге на два: лонги и шорты.

Данная книга является переводом руководства по пакету ***Rattle*** (Простое обучение аналитическим инструментам ***R)*** с примерами для рынка Форекс. Кроме этого книга содержит буквальные и вольные переводы компонент ***R,*** исползуемых ***Rattle***, c комментариями идеологии моделей классификации и результатов их применения.

Книга доступна новичкам, пожелавшим приоткрыть дверь в большой мир предсказания финансовых рынков с помощью моделей классификации, а также полезна опытным трейдерам, желающим быстро, «на скорую руку» проверить свои идеи, при этом получив готовый код для использования в МТ4.

# Содержание

Предисловие 11

О чем? 11

Для кого? 11

Почему *Rattle*? 11

Организация материала 12

Текущее состояние 12

1. Введение 14

1.1. Анализ, прогноз, предсказание 14

1.2. Процесс предсказательного моделирования 14

1.2.1. Разделение данных 14

1.2.2. Целевая переменная 15

1.2.3. Независимые переменные 16

1.2.4. Оценка результативности модели 16

1.2.5. Выбор модели 16

1.2.6. Итоги 16

1.3. Терминология 17

1.4. Используемые наборы данных 19

2. Предварительная обработка данных 20

2.1. Преобразование отдельных предикторов 21

2.1.1. Центрирование и масштабирование 21

2.1.2. Преобразования для исключения асимметрии 21

2.2. Преобразование групп предикторов 22

2.2.1. Преобразования, решающие проблему выбросов 22

2.2.2. Снижение объема данных и выделение предикторов (PCA) 22

2.3. Обработка пропущенных значений 24

2.4. Удаление предикторов 25

2.4.1. Корреляции между предикторами 25

2.5. Добавление предикторов 25

2.6. Группировка предикторов 26

2.7. Функции R 26

3. Переобучение и настройка модели 28

3.1. Проблема переобучения 28

3.2. Настройка модели 28

3.3. Разделение данных 29

3.4. Методы ресемплирования 30

3.4.1. k-кратная кросс-проверка 30

3.4.2. Повторные разделения для обучения/тестирования 31

3.4.3. Бутстрэпинг 32

3.5. Функции R 32

4. Регрессионные модели 34

4.1. Результативность регрессионных моделей 34

4.2. Линейные регрессионные модели 34

4.3. Нелинейные регрессионные модели 35

4.3.1. Нейронные сети 35

4.3.2. Машины опорных векторов (SVM) 36

4.3.3. К-ближайшие соседи 37

4.4. Регрессионные деревья 38

4.5. Бутстрэп агрегированные деревья (bagging) 38

4.6. Случайный лес (random forest) 39

4.7. Усиление (boosting) 40

4.8. Функции R 40

5. Результативность классификационных моделей 42

5.1. Предсказания класса 42

5.2. Основы предсказаний классов 42

6. Линейные классификационные модели 49

6.1. Логистическая регрессия 49

6.2. Линейный дискриминантный анализ (LDA) 49

6.3. Регресия частично наименьших квадратов (PLS) 50

6.4. Функции R 50

7. Нелинейные классификационные модели 51

7.1. Нейронные сети 51

7.2. Машины опорных векторов (SVM) 51

7.3. K-ближайшие соседи (KNN) 52

7.4. Функции R 52

8. Классификационные деревья 54

8.1. Основные классификационные деревья 54

8.2. Бутстрэп агрегированные деревья 55

8.3. Случайные леса 55

8.4. Усиление 55

8.5. Функции R 56

9. Несбалансированность классов 58

9.1. Влияние несбалансированности классов 58

9.2. Настройка модели 58

9.3. Случай неравных весов 58

9.4. Методы сэмплирования 58

9.5. Обучение, чувствительное к стоимости 59

9.6. Функции R 60

10. Значимость предикторов для целевой переменной 61

10.1. Метрики значимости, полученной из моделей 61

10.2. Независимые от модели метрики 62

10.3. Другие подходы 62

10.4. Функции R 62

11. Выбор предикторов 64

11.1. Следствия использования неинформативных предикторов 64

11.2. Подходы для сокращения количества предикторов 64

11.3. Методы обертки 65

11.3.1. Выбор вперед, обратный и пошаговый 66

11.3.2. Генетические алгоритмы 66

11.4. Методы фильтра 66

11.5. Выбор смещения 67

11.6. Инструменты *R* для выбора предикторов 68

11.6.1. Пакет Boruta 68

11.6.2. Пакет varSelRF 68

11.6.3. Пакет FSelector 68

11.6.4. Пакет «CORElearn» 69

11.7. Функции R 69

11.7.1. Выбор вперед, назад и пошаговый 69

11.7.2. Рекурсивное удаление предикторов 69

11.7.3. Методы фильтрации 70

12. Работа новичка с *Rattle* 72

12.1. Интерфейс *Rattle* 72

12.1.1. Проекты 72

12.1.2. Панель инструментов 73

12.1.3. Меню 74

12.1.4. Перемещение с помощью клавиатуры 74

12.2. Введение в моделирование с *Rattle* 74

12.3. Построение модели 74

12.4. Данные (вкладка *Data*) 75

12.5. Исследование (вкладка *Explore*) 77

12.5.1. Опция сводки (Summary) 77

12.5.2. Опция распределения (Distributions) 78

12.5.3. Опция преобразования (Transform) 78

12.5.4. Опция масштабирования (Rescale) 79

12.5.5. Опция заполнения (Impute) 80

12.5.6. Опция перекодирования (Recode) 81

12.6. Модель (вкладка Model) 81

12.6.1. Деревья решений (Tree) 81

12.6.2. Случайный лес (Forest) 83

12.7. Построение всех моделей и их настройки 84

12.7.1. Оценка (вкладка Evaluate) 85

12.7.2. Развертывание модели 86

12.7.3. Скрипты R 86

12.8. Журнал (вкладка Log) 87

12.9. Справка 89

13. Начало работы 91

*13.1.* Стартуем *R* 91

13.2. Выход из *Rattle* и *R* 95

13.3. Пробуем «естественные» котировки 96

13.4. Создание модели на демонстрационных данных погоды 100

13.5. Оценка модели на данных о погоде 105

13.6. Загрузка набора данных с котировками 106

13.7. Описание набора данных с котировками 109

13.7.1. Целевые переменные 111

13.7.2. Котировки 115

13.7.3. Временные метки 117

13.7.4. Сглаженные котировки eurusd 118

13.7.5. Приращения сглаженных котировок eurusd 119

13.7.6. Приращения котировок 119

13.7.7. Приращения котировок через одну 120

13.7.8. Приращения котировок через две 121

13.7.9. Сдвиги вправо сглаженных котировок 122

13.7.10. Индикаторы eurusd 122

14. Вкладка «*Data* - данные» 125

14.1. Номенклатура данных 125

14.2. Данные CSV 127

14.3. Данные ARFF 128

14.4. Источник данных ODBC 128

14.5. Данные из рабочего пространства R 129

14.6. Данные из файлов *R* 130

14.7. Библиотека 132

14.8. Данные корпуса 133

14.9. Опции данных 133

14.9.1. Разделение данных 133

14.9.2. Роли переменных 135

14.9.3. Калькулятор весов и роль 136

14.10. Функции R 136

14.10.1. Пример в R 136

14.10.2. Команда read.table – читать таблицу 138

15. Вкладка «Explore - исследование данных» 140

15.1. Сводки по данным (Summary) 140

15.1.1. Summary/summary 140

15.1.2. Summary/Describe 142

15.1.3. Summary/Basic 142

15.1.4. Summary/Kurtosis - эксцесс 143

15.1.5. Summary/Skewness - асимметрия 143

15.1.6. Summary/Show Missing – показать пропущенные значения 144

15.1.7. Summary/Cross Tab – перекрестная табуляция 144

15.1.8. Функции R 146

15.2. Распределение 148

15.2.1. Ящичковая диаграмма – box plot 150

15.2.2. Гистограмма 153

15.2.3. Кумулятивное распределение 155

15.2.4. Частоты категориальных переменных 156

15.2.5. Функции R 157

15.3. Корреляционный анализ 157

Функции R 161

15.4. Главные компоненты 161

15.4.1. Функции R 164

16. Вкладка «Test - тестирование данных» 170

16.1. Вкладка «Test/ Kolmogorov-Smirnov» 170

16.2. Вкладка «Test/Wilcoxon Rank Sum» 171

16.3. Вкладка «Test/t-test» 172

16.4. Вкладка «Test/F-test» 173

16.5. Вкладка «Test/Correlation» 174

16.6. Вкладка «Test/ Wilcoxon Signed Rank» 174

16.7. Функции R 175

17. Интерактивные графики 176

18. Вкладка «Transform - преобразование данных» 177

18.1. Краткий обзор проблем данных 177

18.1.1. Очистка данных 177

18.1.2. Пропущенные данные 178

18.1.3. Выбросы 178

18.1.4. Выбор переменных 179

18.2. Вкладка «Transform/Rescale – преобразование/масштабирование» 179

18.2.1. Опция «Recenter – центрирование» 180

18.2.2. Опция «Scale[0-1] – масштабирование» 181

18.2.3. Опция «-Median/MAD – медиана» 182

18.2.4. Опции «Natural Log/Log 10 - натуральный логарифм, десятичный логарифм» 182

18.2.5. Опция «Matrix – матрица» 183

18.2.6. Опция «Rank – ранг» 183

18.2.7. Опция «Interval – интервал» 183

18.3. Вкладка «Transform/Impute – преобразование/заполнение» 183

18.3.1. Опция «Zero/Missing - нуль/пропущенные» 184

18.3.2. Средняя/Медиана/Мода 184

18.3.3. Константа 185

18.4. Вкладка «Transform/Recode – перекодирование» 185

18.4.1. Опция «Binning – укладывание в стеллаж» 185

18.4.2. Опция «Indicator Variable – индикативная переменная» 186

18.4.3. Опция «Join categorics - объединение категорий» 187

18.4.4. Опции «As categoric, As numeric - преобразование типов» 188

18.5. Вкладка «Transform/Cleanup – уборка» 188

18.6. Функции R 189

Общий перечень функций 189

19. Описательное и предсказательное моделирование 190

19.1. Терминология моделирования 190

19.2. Платформа для моделирования 190

19.3. Описательное моделирование 192

19.4. Предсказательное моделирование 192

19.5. Модели 193

20. Вкладка «Cluster - кластерный анализ» 195

21. Вкладка «Associate - анализ зависимостей» 197

22. Вкладка «Model/Tree – Модель/дерево решений» 198

22.1. Краткий обзор 198

22.2. Алгоритм 200

22.3. Меры 201

22.3.1. Прирост информации 201

22.3.2. Другие меры 202

22.4. Пример построения дерева 202

22.4.1. Создание модели 202

22.4.2. Древовидная структура 203

22.4.3. Вызов функции 204

22.4.4. Используемые переменные 205

22.4.5. Оценка результативности 205

22.4.6. Предсказание «лонгов-шортов» 206

22.4.7. Отображение деревьев 207

22.5. Настройка параметров 209

22.5.1. Минимальное расщепление (Min Split) 209

22.5.2. Минимальный размер блока (Min Backet) 210

22.5.3. Сложность (Complexity) 210

22.5.4. Предшествующий (Priors) 211

22.5.5. Матрица потерь (Loss Matrix) 212

22.6. Обсуждение 212

22.7. Итоги 212

22.8. Функции R 213

22.8.1. Общий перечень функций 213

22.8.2. Пример в R 213

22.8.3. Функция predict.rpart - предсказание для объекта подгонки Rpart 217

22.8.4. Функция rpart - деревья рекурсивного расщепления и регрессии 218

22.8.5. Функция rpart.control – управление подгонкой rpart 219

23. Вкладка « Model/Forest – Модель/случайный лес» 222

23.1. Краткий обзор 222

23.2. Алгоритм 223

23.2.1. Формирование выборки из набора данных 223

23.2.2. Формирование набора предикторов 223

23.2.3. Случайность 223

23.2.4. Подсчет в ансамбле 224

23.3. Пример построения случайного леса 224

23.3.1. Создание модели 224

23.3.2. Оценка результативности 226

23.3.3. Предсказание «лонгов-шортов» 227

23.3.4. Недостаточно представленные классы 229

23.3.5. Значимость переменных 229

23.3.6. Число переменных 231

23.3.7. Настройка опций 232

23.3.8. Рисунки ошибок 232

23.3.9. Преобразование в правила 232

23.4. Настройка параметров 233

23.4.1. Число деревьев 234

23.4.2. Объем выборки 234

23.4.3. Формирование выборки с заменой 234

23.4.4. Обработка пропущенных значений 234

23.5. Обсуждение 235

23.5.1. Краткая история и альтернативные подходы 235

23.5.2. Использование других случайных лесов 235

23.5.3. Ограничение на категории 235

23.5.4. Меры значимости 236

23.6. Итоги 236

23.7. Функции R 237

23.7.1. Общий перечень функций 237

23.7.2. Пример на R 238

23.7.3. Команда randomForest – классификация и регрессия со случайным лесом 242

*library(randomForest)* 242

Описание 242

23.7.4. Команда rfcv – перекрестная проверка для выбора параметров случайных лесов 245

23.7.5. Команда predict.randomForest – метод предсказания для объектов случайного леса 246

24. Вкладка «Model/Boost – Модель/адаптивное усиление» 249

24.1. Краткий обзор 249

24.2. Алгоритм 249

24.3. Пример построения дерева с адаптивным усилением 251

24.3.1. Создание модели 251

24.3.2. Вызов функции 252

24.3.3. Оценка результативности 253

24.3.4. Предсказание «лонгов-шортов» 255

24.3.5. Количество итераций 257

24.3.6. Используемые переменные 257

24.3.7. Значимость переменных 258

24.3.8. Рисунок ошибок 258

24.3.9. Отображение деревьев 259

24.3.10. Настройка опций 260

24.3.11. Добавление деревьев 261

24.3.12. Потраченное время 261

24.4. Аналоги 261

24.5. Итоги 262

24.6. Функции R 262

24.6.1. Общий перечень функций 262

24.6.2. Пример в R 262

24.6.3. Команда ada – подгонка стохастической модели с усилением 266

24.6.4. Команда predict.ada – предсказание набора данных, используя ada 270

25. Вкладка «Model/SVM – Модель/машина опорных векторов» 272

25.1. Алгоритм 273

25.2. Пример построения машины опорных векторов 274

25.2.1. Создание модели c ядром rbfdot 274

25.2.2. Создание модели c ядром polydot 280

25.2.3. Создание модели c линейным ядром vanilladot 283

25.2.4. Создание модели c ядром tanhdot 285

25.2.5. Создание модели c ядром laplacedot 288

25.2.6. Создание модели c ядром besseldot 290

25.2.7. Создание модели c ядром anovadot 293

25.2.8. Сравнение результативности моделей с разными ядрами 295

25.3. Настройка параметров 295

25.3.1. Тип модели type = 296

25.3.2. Функции ядра kernel = 296

25.3.3. Вероятности класса prob.model = 296

25.3.4. Параметр ядра: стоимость нарушения ограничений C = 296

25.3.5. Параметр ядра: сигма sigma = 296

25.3.6. Перекрестная проверка cross = 296

25.4. Аналоги 296

25.5. Функции R 297

25.5.1. Общий перечень функций 297

25.5.2. Пример в R 297

25.5.3. Функция ksvm - машины опорных векторов 300

25.5.4. Функция predict.ksvm - метод предсказания для объектов «опорный вектор» 304

26. Вкладка «Model/Linear – Модель/линейная» 306

26.1. Краткий обзор 306

26.2. Алгоритм 306

26.3. Вкладка «Model/Linear - Модель/линейная» 307

26.4. Вкладка «Model/Linear/Logistic – Модель/линейная/логистическая» 308

26.4.1. Вызов функции 312

26.4.2. Оценка результативности 313

26.5. Вкладка «Model/Linear/Probit – Модель/линейная/пробит» 315

26.6. Вкладка «Model/Linear/Multinominal - Модель/линейная / многоуровневая» 317

26.7. Функции R 320

26.7.1. Общий перечень функций 320

26.7.2. Пример в R 321

26.7.3. Функция glm(stats) - подгонка обобщенной линейной модели 323

26.7.4. Функция glm.control - дополнение для управления подгонкой GLM 328

27. Вкладка «Model/Neural Net – Модель/нейронные сети» 330

27.1. Краткий обзор 330

27.2. Алгоритм 330

27.3. Пример построения нейронной сети 331

27.3.1. Создание модели 331

27.3.2. Вызов функции 333

27.3.3. Оценка результативности 334

27.3.4. Предсказание «лонгов-шортов» 336

27.4. Функции R 337

27.4.1. Общий перечень функций 337

27.4.2. Пример в R 337

27.4.3. Функция nnet (nnet) - подгонка нейронной сети 340

27.4.4. Функция predict.nnet - предсказание путем использования обученной нейронной сети 342

28. Вкладка «Evaluate - оценить» 343

28.1. Краткий обзор 343

28.2. Меры результативности 344

28.3. Вид владки «Evaluate - оценить» 344

28.4. Error Matrix - матрица ошибок (сопряженности) 346

28.5. Risk - диаграммы риска 349

28.6. Lift - диаграммы лифта 352

28.7. ROC - диаграммы ROC 354

28.8. Sensivity - диаграмма чувствительности/специфики 356

*28.9.* Функции *R* 358

28.9.1. Функция performance -создание объекта performance 358

29. Использование моделей 362

29.1. Введение 362

29.2. Подготовка к использованию моделей 362

29.3. Ручной режим 362

29.3.1. Подготовка исходных данных для торговли 362

29.3.2. Подготовка модели к использованию 363

29.3.3. Пример использования моделей в R 365

29.4. Автоматизированный режим (терминал МТ4) 365

29.4.1. Организация обмена между терминалом MetaTrader 4 и R 366

29.4.2. Оформление модели случайного леса 367

29.4.3. Организация обмена данными между советником и моделью 367

29.4.4. Обмен данными между терминалом МТ4 и моделью по приходу каждого бара 368

Приложение А. Установка *Rattle* 370

Приложение В. Наборы данных 373

B1. Набор данных *audit* 373

B2. Набор данных *weather* 373

B3. Набор данных kot60\_110101\_131231\_UE.txt 373

B4. Набор данных *zz\_1\_5.RData* 374

Приложение С. Тексты скриптов 375

C1. *Prepare\_DF.R* - cкрипт подготовки исходных данных 375

C2. *prep\_zz* - функция подготовки файла обучения и предсказания 376

C3. prep\_model – скрипт сохранения модели 385

C4. *Predict\_model* - cкрипт предсказания (без МТ4) 386

C5. *predict\_rf* - функция предсказания по модели случайного леса (из МТ4) 387

Литература 389

# Предисловие

## О чем?

При построении торговых систем вообще, и с использованием терминала МТ4/5 в частности, приходится решать целый комплекс взаимосвязанных задач.

Изначально, целью построения торговой системы является *предсказание* поведения некоторого рыночного инструмента, например, валютной пары. Цели предсказания могут быть разными, мы же ограничимся предсказанием *трендов*, а точнее предсказанием *роста (лонгов)* или *падения («шортов»)* значений котировки валютной пары. Кроме этого будем предсказывать боковики - нахождение *вне* рынка.

## Для кого?

Книга доступна для многих читателей и не обязательно только тем, кто силен в информатике или статистике. С помощью ***Rattle*** практически любой желающий сможет построить основную часть торговой системы – предсказание котировки, а затем, при отсутствии необходимого опыта и знаний, сделать заказ реальной торговой системы, изложив свои мысли в виде готового кода на ***R.***

Для искушенного в ***R*** пользователя ***Rattle*** будет также полезен: позволяет быстро апробировать идеи в исходных данных, целевых переменных, используемых моделях, а затем перейти к соответствующим пакетам ***R,*** имеющим значительно больший по сравнению с ***Rattle*** функционал.

Данная книга является руководством по использованию пакета ***Rattle*** (Простое обучение аналитическим инструментам ***R)***, который оформлен в виде GUI – графического пользовательского интерфейса, позволяющего значительно упростить использование могучих средств ***R*** и необходимых для поставленной задачи пакетов.

## Почему *Rattle*?

В качестве инструмента для предсказания поведения валютных пар выберем систему ***R,*** которая идеально подходит для задач предсказания на финансовых рынках и, в частности, предсказания поведения валютных пар. Вместе с тем ***R*** остается, прежде всего, языком программирования для высоко квалифицированного статистика и для многих остается вне досягаемости. Сложность самой ***R*** усугубляется тем обстоятельством, что инструменты для предсказания являются многочисленными и рассредоточены по многим *пакетам*, которые и составляют основную функциональность ***R.***

***Rattle*** объединяет множество пакетов ***R,*** которые важны для построения торговых систем, но часто не легки для использования новичком. Понимание ***R*** не требуется, чтобы начать с ***Rattle***. Но результатом работы с ***Rattle*** будет код на ***R,*** который может быть использован при построении реальной торговой системы. И на этом этапе потребуется знание ***R.***

В любом случае ***Rattle*** является незаменимым инструментом на этапе проектирования торговой системы, позволяет даже новичкам быстро посмотреть результат тех или иных идей и получить их оценку.

***Rattle*** (Уильямс, 2009) является бесплатным программным обеспечением с открытым исходным кодом, созданном в рамках статистического пакета программного обеспечения ***R*** (R Рабочая группа Разработки, 2011). Как бесплатное программное обеспечение исходный код ***Rattle*** и ***R*** доступен всем без ограничения. Исходный код ***Rattle*** написан на ***С*** и всем разрешено, и действительно поощряется, рассмотрение исходного кода для обучения, его понимания, проверки и расширения.

## Организация материала

Книга состоит из следующих частей.

**Часть 1.** Введение в предсказательное моделирование изложено в главах 1-11. В этой части книги сжато, но достаточно подробно на описательном уровне рассматриваются основные понятия предсказательного моделирования. Необходимость этой части проистекает из того, что авторы ***Rattle*** не объясняют смысл и взаимодействие различных частей своей системы. Изучение первой части позволит осмысленно подобрать в ***Rattle*** инструменты для решения конкретно Вашей задачи.

**Часть 2.** Краткое описание ***Rattle*** изложено в главе 12. Эта часть полезна как на этапе первоначального знакомства с ***Rattle***, так и на этапе постоянного использования в качестве краткого справочника.

**Часть 3**. Полное описание ***Ratlle*** 13-29. Эта часть книги представляет собой перевод руководства по ***Rattle***. К авторскому руководству добавлены примеры для рынка Форекс, а также приведены переводы синтаксиса команд ***R,*** которые использует ***Rattle***.

## Текущее состояние

Новые версии ***R*** публикуются два раза в год - в апреле и октябре. ***R*** имеет несколько миллионов пользователей, что гарантирует очень высокий уровень отлаженности кода. Система статистики и графики ***R*** очень популярна, быстро расширяется за счет пакетов, имеет обширную информационную поддержку в виде публикаций, учебников и монографий.

Примеры, включенные в эту книгу, используют версию 3.1.1 ***R*** и версию 3.0.2 r169 ***Rattle***. ***Rattle*** - развивающийся пакет и, хотя понятия остаются, детали меняются. Поэтому не следует удивляться, если скриншоты, приведенные в этой книге, будут отличаться от Ваших скриншотов.

Часть 1. Введение в предсказательное моделирование

Первая часть книги является введением в идеологию предсказательного моделирования. В этой части книги сжато, но достаточно подробно на описательном уровне рассматриваются основные понятия предсказательного моделирования.

Первая часть книги дополняет техническую документацию по ***Rattle,*** так как авторы ***Rattle*** предполагают, что пользователь их продукта знаком с терминологией, проблемами и инструментами, существующими в области предсказательного моделирования.

Первая часть книги будет полезна всем без исключения проектировщиками предсказательных моделей вообще, а не только пользователям ***Rattle.*** Излагаемые в первой части материалы охватывают более широкий круг вопросов, чем необходимо для работы с ***Rattle,***готовя читателя к использованию других, аналогичных, но более развитых инструментов для построения предсказательных моделей.

# Введение

## Анализ, прогноз, предсказание

На финансовых рынках будем различать следующий набор действий: анализ, прогноз и предсказание.

**Анализ** позволяет ответить нам на вопрос: почему так произошло? Например, можно поставить вопрос: почему произошло падение курса доллара по отношению к евро? Без анализа прошлого, без анализа исторических данных невозможен переход к последующим этапам – *прогнозу* или *предсказани*ю.

**Прогноз.** Значение слова «прогноз» буем понимать так, как это понимается в ***R*** под словом «*forecast*» - для прогноза следующего значения используется предыдущее значение, полученное в результате предыдущего шага прогноза. Пакет *forecast* является примером такого понимания значения слова «прогноз».

**Предсказание.** Значение слова *«предсказание»* будем понимать в смысле универсальной функции *predict –* предсказание будущего на любое число шагов вперед с использованием имеющихся данных.

В данной книге исторические данные используются для **обучения** моделей, которые в последующем используются для **предсказания** будущего**.**

## Процесс предсказательного моделирования

Существует несколько аспектов в процессе построения модели, которые следует обсудить далее, особенно новичкам в предсказательном моделировании.

Технология предсказательного моделирования выглядит следующим образом:

* на основе некоторого набора исходных данных производится обучение *модели*;
* в последующем обученная модель используется для предсказания целевой переменной на новом наборе данных;
* в зависимости от того, чему мы учили модель: предсказывать целевую переменную на сегодня, на завтра или на ***n*** шагов вперед, мы и получим соответствующее предсказание.

### Разделение данных

То, как выделяются данные определенным этапам (например, обучение модели, оценка результативности), является важным аспектом моделирования. Наш главный интерес, к примеру, состоит в предсказании тренда на новых данных, которые отсутствуют в момент обучения модели. Это означает, что до некоторой степени необходимо проверить, как хорошо модель *экстраполирует* на новых котировках. Если бы мы интересовались предсказанием на тех же самых данных (то есть, *интерполяцией*), то можно было бы взять простую случайную выборку данных. То, как определены набор данных для обучения и наборы данных для тестирования и проверки, должно отразить применение модели.

Сколько данных должно быть выделено обучающему набору и тестовому? Это обычно зависит от ситуации. Если пул данных небольшой, решения разделения данных могут быть критическими. Небольшой (десятки наблюдений) тестовый набор исходных данных ограничил бы суждения о результативности модели. В этом случае, уверенность в результатах могли бы предоставить методы *ресемплирования,* которые могли бы решить проблему.

На финансовых рынках достаточно часто доступны большие наборы данных. Обычно доступны наборы данных с несколькими тысячами наблюдений, например, цен валютных пар. Казалось бы, что на финансовых рынках отсутствуют проблемы небольших по объему исходных данных. Однако это не так. К примеру, мы обучили модель на растущем рынке, а тестируем на падающем рынке, и выясняется, что модель дает убыток на этом тестовом наборе данных. На этом примере мы видим, что проблема в обязательности разделения котировок для обучения и тестирования остается и для больших наборов данных, так как при разделении необходимо обеспечить примерно равное количество «лонгов» и «шортов».

### Целевая переменная

Несмотря на кажущуюся простоту, выбор цели предсказания и данных, которые материализуют эту цель в виде набора цифр, носит краеугольный характер.

Обратимся к идее предсказывать тренд. Эта идея опирается на желание торговать тренды.

Из определения тренда: «тренд растущий, если следующая цена больше предыдущей цены» и, наоборот, для падающего тренда. Из определения следует, что необходимо предсказывать цену валютной пары. Было 1.3500 для *eurusd*, предсказали 1.3550 – растущий тренд, покупаем.

Но базовыми приказами являются приказы «купить» и «продать», а мы предсказали *величину* цены. А величина цены используется, к примеру, в торговых системах на пробой уровня. Чтобы реализовать план по торговле трендов, надо будет произвести дополнительное действие по сравнению цен. При этом очевидно, что мы предсказываем не то, что собирались торговать!

Поэтому, если мы собрались делать трендовую торговую систему, то модель должна сразу предсказывать тренд. Учить модель надо трендам, целевая переменная должна принимать только два значения «купить» и «продать» или в закодированном (категориальном) виде «1» и «-1».

Можно уточнить целевую переменную, имеющую два значения за счет того, что неплохо бы предсказывать и боковики, т.е. иметь целевую переменную, принимающую значения *«купить», «продать»* и «*вне рынка».* Или закодировав эти значения как: «1», «-1» и «0».

Различие моделей, которые используют совокупность исходных данных для вычисления будущей *величины* цены финансового актива, и моделей, которые относят совокупность исходных данных к некоторому классу, является принципиальным. Первый тип моделей относится к *регрессионным моделям*, а второй тип моделей относится к *классификационным моделям*.

По предсказательным моделям регрессионного типа вычисляется значение некоторой величины в будущем, и когда это будущее наступит, то у нас будет фактическое значение этой предсказанной величины.

По предсказательным моделям классификационного типа вычисляется класс, к которому будет отнесена совокупность поступивших на момент предсказания исходных данных.

Рассмотренные варианты не исчерпывают всего разнообразия целевых переменных, возможных на финансовых рынках. Но вывод из данного раздела: целевая переменная должно точно соответствовать целям торговой системы.

### Независимые переменные

Независимые переменные, в дальнейшем *предикторы*, независимы в том смысле, что поступают в модель извне, являются внешними, измеряемыми переменными, или переменными, вычисленными на основе этих внешних переменных. Например, любые экономические, финансовые данные, включая котировки валютных пар, являются независимыми переменными, так как их значения образуются в результате деятельности субъектов на рынке. К этой же категории переменных относятся и индикаторы из технического анализа, которые вычисляются на основе котировок.

Выбор независимых переменных не менее важен, чем выбор целевой переменной. Более того, именно выбор независимых переменных определяет успешность моделирования. Основное время, затраченное на разработку модели, уходит как раз на анализ и подбор набора независимых переменных.

Этот вопрос рассмотрим в отдельных разделах.

### Оценка результативности модели

Тип модели предполагает разные типы оценок.

Для регрессионных моделей – это ошибка предсказания, полученная как разность между предсказанной и фактической величиной (к примеру, RMSE).

Для классификационных моделей – это рассогласование, полученное как совпадение/несовпадение фактических и предсказанных классов.

### Выбор модели

Наличие оценки результативности модели позволяет выбрать лучшую модель. Это можно сделать, если «лучшая» модель сильно отличается от своих конкурентов. Если это не так, то, отбросив «худшую» модель при ее наличии, можно сделать предсказание, используя имеющиеся предсказания моделей в качестве предикторов для окончательного предсказания.

### Итоги

В первом приближении создание модели кажется ясным: выбираем метод моделирования, учим модель на наборе данных обучения – все готово, можно предсказывать.

Вам очень повезет, если столь просто удастся создать надежную, устойчивую модель, работающую на новых наблюдениях.

Чтобы получить модель, имеющую примерно одинаковые оценки результативности вне набора обучения следует сначала понять данные и цель моделирования. После понимания данных и целей, можно предварительно обработать и разделить данные. После этих шагов можно начинать создание, оценку и выбор моделей. Только после того, как эти шаги сделаны, мы, наконец, начнем создание, оценку и выбор моделей.

Существует целый ряд общих причин неудачности предсказательных моделей:

* не адекватная предварительная обработка данных;
* не адекватная проверка модели;
* неоправданная экстраполяция (применение к данным, которые имеют слабое отношение к обучающему набору);
* **наиболее важное**: *переобучение* модели на обучающем наборе данных.

## Терминология

Предсказательное моделирование является одним из многих наименований, которые относятся к процессу выявления отношений внутри данных для предсказания желаемого результата. Машинное обучение, искусственный интеллект, распознавание образов, интеллектуальный анализ данных, предсказательная аналитика - много научных областей сделало вклад, что привело к синонимии разных понятий.

*Предсказательное моделир*ование – это процесс, с помощью которого модель создает, выбирает или пытается сделать лучшее предсказание вероятности результата.

*Набор данных* – это общий и расплывчатый термин.

Набор данных на внешнем носителе – это файл данных по тексту книги. По расширению файла можно судить о кодировке и, частично, о структуре файла. В ***Rattle*** допустимы разные файлы. Наибольший интерес для нас будут представлять файлы со следующими расширениями:

* *.txt* – обычный текстовый файл;
* *.csv* – текстовый файл Excel;
* *.RData* – файл ***R,*** в котором хранится рабочая область.

Набор данных в памяти – это некоторая совокупность данных, имеющая структуру. В терминах ***R*** – это вектор, матрица, фрейм данных или совокупность этих данных.

Матрица (редко) и фрейм данных в ***Rattle*** представлены таблицей, имеющей следующий вид:

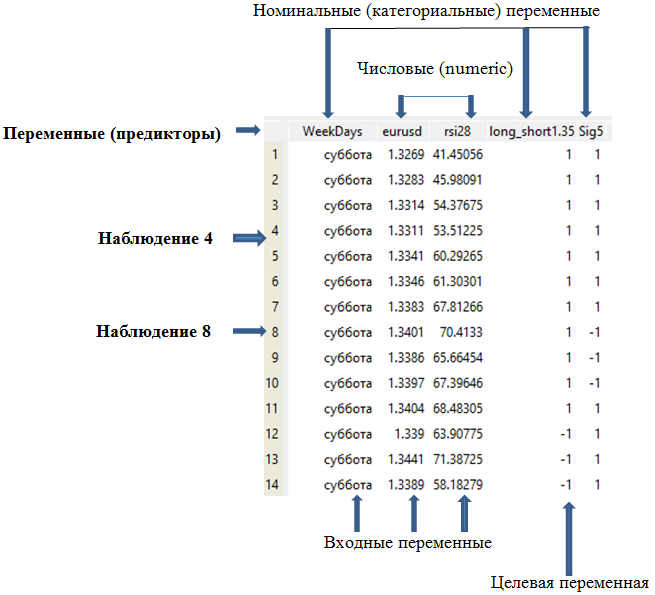


Рис.1.1. Фрейм данных, представленный в ***Rattle***

Термины *выборка (sample), наблюдение (observation), пример, экземпляр (instance)* относится к отдельной строке данных. Термин *sample* также может относить к подмножеству наблюдений, которые объединены, например целью последующего использования – *обучающая* выборка или *обучающий* набор данных. Значение термина выборка будет понятно из контекста употребления термина.

*Обучающий набор* содержит данные, которые использовались для обучения модели, в то время как *тестовый* и *проверочный* наборы используются исключительно для оценки результативности модели.

*Предикторы*, *независимые переменные, атрибуты или дескрипторы* являются данными, которые используются в качестве входных переменных в уравнении предсказания. На рис.1.1 показаны три предиктора, которые играют роль в модели «входных переменных».

*Результат, зависимая переменная, целевая переменная, класс, отклик* (response) относится к результирующему событию или количеству, которое предсказывается.

У *числовой переменной* есть значение, которое является целым числом или вещественным числом, такими как цена валютной пары, объем торгов, процентная ставка. Числовые переменные также известны как *количественные переменные*. Числовые переменные могут быть дискретными (целыми числами) или непрерывными (действительными). Например, котировка валютной пары. У числовой переменной обычно имеется числовой масштаб. Для валютной пары *eurusd* числовой масштаб – это диапазон от 0.5 до 2.0, в который укладываются все имевшие место значения цен на эту валютную пару. Совершенно другой масштаб у валютной пары *usdjpy* – величины цен на эту валютную пару почти на два порядка больше, чем на *eurusd.*

*Категориальные* (*categorical*) данные, известные также как номинальные атрибуты, качественные данные, факторы имеют значения, которые не имеют масштаба. «Лонг/шорт», день недели являются примерами таких данных. «Лонг» не больше и не меньше «шорта». Категориальная переменная, которая имеет два значения, как у нас – (лонг, шорт) называют *бинарной* (двоичной) переменной. Категориальная переменная «день недели» имеет семь значений.

Категориальные переменные могут быть упорядочены, как в нашем примере *Weekdays* (дни недели). Понедельник не больше и не меньше вторника, но может быть важным для модели, чтобы ей было известно, что вторник всегда следует после понедельника.

*Построение модели, обучение модели, тренировка модели или оценка параметров* – все это относится к процессу определению параметров в уравнении модели.

## Используемые наборы данных

Далее по тексту будут использоваться следующие наборы данных:

*audit* набор данных, поставляемый в составе дистрибутива ***Rattle***.

*weather* набор данных, поставляемый в составе дистрибутива ***Rattle***.

*kot60\_110101\_131231\_UE.txt*

На основе регрессионной модели попытаемся сделать «типичный мультивалютник»:

* целевая переменная – *EURUSD;*
* предикторы – *GBPUSD, USDCHF,USDJPY, EURGBP, USDCAD.*

*zz\_1\_5.RData*

Для классификационной модели создан искусственный разнообразный набор предикторов, которые должны продемонстрировать возможности моделей по предсказанию трендов:

* целевая переменная (три варианта) – тренд;
* предикторы – день недели, час дня, *EURUSD*, приращение *EURUSD*, *macd, macd(13), macd(26), macd(39*), приращение *macd(13), macd(26), macd(39)), RSI(14), RSI(21), RSI(28),* стелажирование на 8 уровней *(RSI(14), RSI(21), RSI(28)), MA(13), MA(26), MA(52)*, приращение (*MA(13), MA(26), MA(52)).*

Описание каждого набора данных приведено в Приложении В. Для *zz\_1\_5.RData* приведен скрипт на ***R,*** который формирует этот набор из набора *kot60\_110101\_131231\_UE.txt*

