Машинное обучение

Хенрик Бринк
Джозеф Ричардс
Марк Феверолф

MANNING

ББК 32.973.236
УДК 004.85
Б87

Бринк Хенрик, Ричардс Джозеф, Феверолф Марк

Б87 Машинное обучение. — СПб.: Питер, 2017. — 336 с.: ил. — (Серия «Библиотека программиста»).

ISBN 978-5-496-02989-6

В последние годы машинное обучение вышло на уровень большого бизнеса: компании активно используют его для зарабатывания денег, прикладные исследования бурно развиваются, а неугомонные разработчики используют любую возможность повысить свой уровень владения этой тематикой.

Данная книга рассчитана на тех, кто хочет решать самые разнообразные задачи при помощи машинного обучения. Как правило, для этого нужен Python, поэтому в примерах кода используется этот язык, а также библиотеки pandas и scikit-learn. Вы познакомитесь с основными понятиями ML, такими как сбор данных, моделирование, классификация и регрессия, а главное, получите практический опыт обработки реальных данных.

12+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.973.236
УДК 004.85

Права на издание получены по соглашению с Manning. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги.

ISBN 978-1617291920 англ.
ISBN 978-5-496-02989-6

© 2017 by Manning Publications Co. All rights reserved
© Перевод на русский язык ООО Издательство «Питер», 2017
© Издание на русском языке, оформление ООО Издательство «Питер», 2017
© Серия «Библиотека программиста», 2017

Оглавление

Предисловие **13**

Вступление **15**

Благодарности **17**

О книге **18**

Структура книги 18

Как читать эту книгу 19

Целевая аудитория 20

Формат кода, загрузки и требования к ПО 20

От издательства 21

Об авторах **22**

Часть I. Последовательность действий при машинном обучении **23**

Глава 1. Что такое машинное обучение? **24**

1.1. Как обучаются машины 25

1.2. Принятие решений на основе данных 30

1.2.1. Традиционные подходы 32

1.2.2. Подход с машинным обучением 36

1.2.3. Пять преимуществ машинного обучения 42

1.2.4. Сложности 43

1.3. Рабочий процесс: от данных до внедрения 44
1.3.1. Сбор и подготовка данных 45
1.3.2. Обучение модели на данных 46
1.3.3. Оценка производительности модели 49
1.3.4. Оптимизация производительности модели 50
1.4. Усовершенствованные способы повышения эффективности 51
1.4.1. Предварительная обработка данных и проектирование признаков 51
1.4.2. Непрерывное совершенствование моделей 54
1.4.3. Масштабирование моделей 54
1.5. Заключение 55
1.6. Терминология 56

Глава 2. Реальные данные 57

2.1. Первый этап: сбор данных 59
2.1.1. Определяем набор входных признаков 62
2.1.2. Наблюдаемое значение целевой переменной 64
2.1.3. Достаточный объем обучающих данных 65
2.1.4. Репрезентативность обучающей выборки 68
2.2. Подготовка данных к моделированию 69
2.2.1. Категориальные признаки 70
2.2.2. Отсутствующие данные 73
2.2.3. Основы проектирования признаков 76
2.2.4. Нормализация данных 78
2.3. Визуализация данных 80
2.3.1. Мозаичные диаграммы 81
2.3.2. Диаграммы размаха 83
2.3.3. Графики плотности 86
2.3.4. Диаграммы рассеяния 88

2.4. Заключение 89
2.5. Терминология 90

Глава 3. Моделирование и прогнозирование 91

3.1. Основы моделирования с машинным обучением 92
3.1.1. Поиск связи между входными данными и целевой переменной 93
3.1.2. Зачем нужна хорошая модель 95
3.1.3. Типы методов моделирования 97
3.1.4. Обучение с учителем и без 100
3.2. Классификация: распределение по классам 101
3.2.1. Построение классификатора и получение предсказаний. . . 103
3.2.2. Классификация сложных нелинейных данных 108
3.2.3. Классификация в случае множества классов 111
3.3. Регрессия: предсказание численных значений 113
3.3.1. Построение регрессора и генерация прогнозов 115
3.3.2. Регрессия для сложных нелинейных данных 119
3.4. Заключение 121
3.5. Терминология 122

Глава 4. Оценка и оптимизация модели 123

4.1. Оценка прогностической точности на новых данных 125
4.1.1. Проблема: переобучение и чрезмерно оптимистическая оценка модели 125
4.1.2. Решение: скользящий контроль 129
4.1.3. На что следует обращать внимание при перекрестной проверке. 134
4.2. Оценка моделей классификации 135
4.2.1. Точность для отдельных классов и таблица сопряженности 138

4.2.2. Компромиссы при оценке точности и ROC-кривые 140
4.2.3. Многоклассовая классификация 144
4.3. Оценка моделей регрессии 147
4.3.1. Простые показатели эффективности регрессионных моделей 148
4.3.2. Исследование остатков 150
4.4. Оптимизация модели путем подбора параметров 152
4.4.1. Параметры настройки ML-алгоритмов 152
4.4.2. Сеточный поиск 154
4.5. Заключение 158
4.6. Терминология 159

Глава 5. Основы проектирования признаков 161

5.1. Мотивация: в чем польза проектирования признаков? 162
5.1.1. Что такое проектирование признаков? 162
5.1.2. Пять причин проектирования признаков 163
5.1.3. Проектирование признаков и знание предметной области 165
5.2. Основные этапы проектирования признаков 166
5.2.1. Пример: рекомендация события 167
5.2.2. Обработка даты и времени 170
5.2.3. Извлечение признаков из обычного текста 172
5.3. Выбор признаков 174
5.3.1. Прямой отбор и обратное исключение 178
5.3.2. Отбор признаков для исследования данных 180
5.3.3. Практический пример отбора признаков 181
5.4. Заключение 184
5.5. Терминология 186

Часть II. Практическое применение 187

Глава 6. Пример: чаевые для таксистов 188

6.1. Данные: сведения о чаевых и плате за проезд 189

6.1.1. Визуализация данных 190

6.1.2. Формулировка задачи и подготовка данных 194

6.2. Моделирование 197

6.2.1. Базовая линейная модель 197

6.2.2. Нелинейный классификатор. 199

6.2.3. Добавление категориальных признаков 202

6.2.4. Добавление временных признаков 203

6.2.5. Аналитическая оценка модели 205

6.3. Заключение 206

6.4. Терминология 207

Глава 7. Усовершенствованное проектирование признаков . . . 208

7.1. Более сложные текстовые признаки 209

7.1.1. Модель «мешок слов» 209

7.1.2. Тематическое моделирование 213

7.1.3. Расширение содержимого 217

7.2. Признаки, извлекаемые из изображений 219

7.2.1. Простые признаки 220

7.2.2. Извлечение объектов и форм 222

7.3. Признаки временных рядов 228

7.3.1. Типы временных рядов 228

7.3.2. Предсказания на основе временных рядов 231

7.3.3. Признаки классических временных рядов. 232

7.3.4. Проектирование признаков для потоков событий 238

7.4. Заключение 239
7.5. Терминология 241

Глава 8. Пример обработки естественного языка 243

8.1. Изучение данных и сценарии их применения 244
8.1.1. Первый взгляд на набор данных 245
8.1.2. Анализ набора данных. 246
8.1.3. Так какой же будет наша задача? 247
8.2. Генерация базовых NLP-признаков и построение первого варианта модели 252
8.2.1. Признаки из «мешка слов» 253
8.2.2. Модель на базе наивного байесовского классификатора. . . 255
8.2.3. Нормализация признаков, полученных из «мешка слов», алгоритмом tf-idf 260
8.2.4. Оптимизация параметров модели. 262
8.3. Усовершенствованные алгоритмы и тонкости процесса внедрения 267
8.3.1. Word2vec-признаки 268
8.3.2. Модель на базе алгоритма «случайный лес» 270
8.4. Заключение 273
8.5. Терминология 274

Глава 9. Масштабирование процесса машинного обучения 275

9.1. Перед началом масштабирования 276
9.1.1. Определяем важные аспекты. 277
9.1.2. Прореживание обучающей выборки вместо масштабирования? 280
9.1.3. Масштабируемые системы управления данными. 282
9.2. Масштабирование конвейера ML-моделирования 285
9.2.1. Масштабирование обучающих алгоритмов 286

9.3. Масштабирование предсказаний 291
9.3.1. Масштабирование объема предсказаний 292
9.3.2. Масштабирование скорости предсказаний 293
9.4. Заключение 296
9.5. Терминология 298

Глава 10. Пример с цифровой рекламой 300

10.1. Показ рекламы 302
10.2. Данные, связанные с цифровой рекламой 303
10.3. Проектирование признаков и стратегия моделирования 304
10.4. Размер и форма данных 306
10.5. Сингулярное разложение 309
10.6. Оценка и оптимизация ресурсов 312
10.7. Моделирование 314
10.8. Метод k-ближайших соседей 315
10.9. «Случайные леса» 318
10.10. Другие практические моменты 319
10.11. Заключение 321
10.12. Терминология 322
10.13. Подводим итоги 323

Приложение. Популярные алгоритмы машинного обучения . . . 326

1 Что такое машинное обучение?

В этой главе:

- ✓ основы машинного обучения;
- ✓ преимущества машинного обучения перед традиционными методами;
- ✓ базовые этапы машинного обучения;
- ✓ усовершенствованные методы повышения эффективности моделей.

В 1959 г. специалист по вычислительной технике из компании IBM Артур Самуэль написал компьютерную программу для игры в шашки. Каждому положению на доске присваивался некий вес, базирующийся на вероятности выигрыша. Изначально вероятность определялась по формуле, в которой учитывались такие факторы, как количество шашек на каждой стороне и количество дамок. Подход работал, но Самуэль придумал, каким образом можно повысить его эффективность. Сыграв с программой тысячу партий, он использовал их результаты для уточнения позиционных весов. К середине 1970-х гг. программа достигла уровня хорошо подготовленного непрофессионального игрока.[1]

Самуэль написал компьютерную программу, которая могла по мере накопления опыта улучшать собственные результаты. Программа училась — так зародилось машинное обучение (ML — machine learning).

Мы не собираемся вдаваться в запутанные и сложные математические подробности алгоритмов машинного обучения (хотя и «снимем несколько верхних листьев с этого капустного кочана», чтобы дать вам представление о функционировании наиболее распространенных алгоритмов). Но, по сути, основной целью книги является предоставление неспециалистам информации о важных аспектах и распространенных проблемах, с которыми приходится сталкиваться при интеграции машинного обучения в приложения и конвейеры данных. В этой главе мы рассмотрим реальную экономическую задачу — обзор кредитных заявок, которая покажет преимущество машинного обучения перед большинством существующих альтернатив.

1.1. Как обучаются машины

У людей мы различаем механическое заучивание и интеллектуальное осмысление. Зазубривание телефонных номеров или инструкций, без сомнения, тоже относится к процессу *обучения*. Но, как правило, под этим понятием мы подразумеваем кое-что другое.

Ребенок, играющий с друзьями, наблюдает реакцию других членов группы на свои действия. Этот опыт влияет на его будущее поведение

[1] Jonathan Schaeffer. *One Jump Ahead: Computer Perfection at Checkers* (New York: Springer, 2009).

в социуме. Но он не вспоминает и не проигрывает заново свое прошлое, а опирается на определенные, легко опознаваемые характеристики прошлых взаимодействий: детская площадка, класс, мама, папа, сестры и братья, друзья, незнакомцы, взрослые, дети, в помещении или на улице. Оценка новой ситуации базируется на признаках, с которыми ему доводилось сталкиваться раньше. Обучение при этом является не просто сбором информации. Формируется то, что можно назвать *аналитической оценкой*.

Представьте, как вы по картинкам учите ребенка отличать собаку от кошки. Показанная картинка кладется в одну из двух стопок, в зависимости от правильности полученного ответа. Чем дольше продолжается процесс, тем выше эффективность распознавания. Что интересно, нет необходимости специально учить ребенка отличать собаку от кошки. Человеческое сознание обладает встроенными механизмами классификации. Ему требуются только *образцы*. Научившись работать с картинками, ребенок сможет опознать практически любое изображение кошки или собаки, не говоря уже о реальных животных. Эта способность *обобщать*, применяя полученные в процессе тренировок знания к новым, ранее не встречавшимся образцам, является ключевой характеристикой как человеческого, так и машинного обучения.

Разумеется, процесс получения знаний человеком превосходит своей сложностью самые совершенные алгоритмы машинного обучения, но у компьютера есть преимущество в виде большей емкости для запоминания, извлечения и обработки данных. Накапливаемый им опыт представлен в форме данных за длительный период времени, обработанных с помощью описанных в этой книге техник, причем это представление позволяет получать и оптимизировать алгоритмы, реализующие если не аналитическую оценку, то хотя бы способность к обобщениям.

Аналогия между человеческим и машинным обучением закономерно заставляет вспомнить такое явление, как *искусственный интеллект* (AI — artificial intelligence). При этом естественным образом возникает вопрос: «Чем искусственный интеллект отличается от машинного обучения?». По этому вопросу нет единого мнения, но большинство соглашается с тем, что ML — это одна из форм AI, так как AI представляет собой куда более обширную область, включающую в числе прочего робототехнику, обработку лингвистической информации и системы машинного зрения. Неоднозначность терминологии усиливает тот факт, что машинное обу-

чение все чаще применяется во многих сопутствующих областях AI. Можно сказать, что такая дисциплина, как машинное обучение, относится к *специализированной совокупности знаний и связанным с ней техникам*. Легко определить, что относится, а что не относится к машинному обучению, в то время как в случае искусственного интеллекта далеко не всегда можно провести такую же четкую границу. Перефразируя часто цитируемое определение Тома Митчелла, скажем, что компьютерная программа обучается, если ее производительность при выполнении определенной задачи, выраженная в измеряемых единицах, увеличивается по мере накопления опыта.[1]

Компания Kaggle объявила конкурс на алгоритм, максимально точно отличающий собак от кошек.[2] Для тренировки участникам предоставили 25 000 изображений с метками, указывающими, кто именно изображен на картинке. После обучения каждый алгоритм должен был классифицировать 12 500 не имеющих меток тестовых изображений.

Те, кому мы рассказывали об этом конкурсе, зачастую задумывались о признаках, по которым можно отличить собаку от кошки. У кошек треугольные и стоячие уши, а у собак они висят, но бывают и исключения. Представьте, что вы должны, не прибегая к иллюстрациям, объяснить разницу между кошкой и собакой человеку, который никогда не видел ни того ни другого животного.

Для обучения и обобщения люди используют различные данные из примеров, включая формы, цвета, текстуры, пропорции и другие характеристики. Машинное обучение также применяет множество стратегий в различных комбинациях в зависимости от поставленной задачи.

Эти стратегии нашли свое воплощение в наборе алгоритмов, разработанных в течение последних десятилетий как учеными, так и практиками в самых разных дисциплинах — от статистики, компьютерной науки, робототехники и прикладной математики до поиска в Интернете, развлекательной сферы, цифровой рекламы и переводов с одного языка на другой. Алгоритмы крайне разнообразны и имеют свои сильные

---

[1] Tom Mitchell, *Machine Learning* (McGraw Hill, 1997): «Говорят, что компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта E».

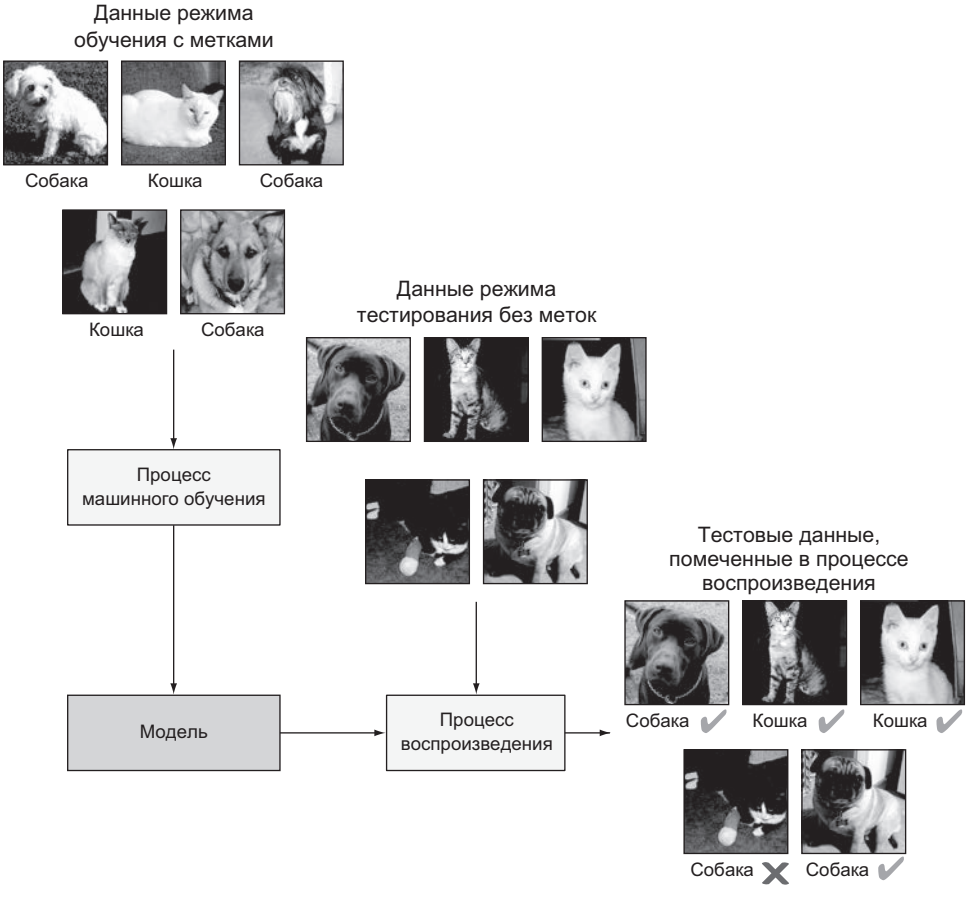
[2] См. страницу «Собаки против кошек» на сайте *www.kaggle.com/c/dogs-vs-cats*.

и слабые стороны. Некоторые относят объекты к определенному классу, другие предсказывают числовые значения. Существуют и алгоритмы, определяющие сходства и различия допускающих сравнение сущностей (например, людей, машин, процессов, кошек, собак). При этом все алгоритмы обучаются на примерах (опыте) и умеют применять полученные знания к новым, ранее не встречавшимся случаям, то есть способны к обобщению.

Заявленные на конкурс «Собаки против кошек» программы на этапе обучения раз за разом пытались корректно выполнить классификацию, используя множество алгоритмов. На каждой из миллионов обучающих итераций программа производила классификацию, измеряла полученный результат и затем хотя бы немного корректировала процесс в поисках постепенного улучшения. Победитель смог корректно распознать 98,914% ранее не демонстрировавшихся тестовых изображений. Это замечательный результат, если учесть, что у людей частота появления ошибки составляет примерно 7%. Процедура показана на илл. 1.1. Процесс машинного обучения анализирует изображения с метками и строит модель, которая в свою очередь используется процессом *воспроизведения* (предсказания) для классификации новых изображений. В примере вы видите, что одно изображение с кошкой было распознано некорректно.

Обратите внимание, что в данном случае мы рассматриваем так называемое *обучение с учителем* (supervised machine learning), но существуют и другие типы машинного обучения. Позже мы поговорим и о них.

Машинное обучение применяется к широкому кругу экономических задач — от обнаружения мошенничества до выбора целевой аудитории и рекомендаций товара, наблюдения за производством в реальном времени, анализа тональности текстов и медицинской диагностики. Оно может взять на себя задачи, которые невозможно выполнить вручную из-за огромного количества подлежащих обработке данных. В случае больших наборов данных машинное обучение иногда обнаруживает неочевидные зависимости, которые невозможно распознать при сколь угодно скрупулезном ручном рассмотрении. При этом комбинация множества таких «слабых» соотношений дает прекрасно работающие механизмы прогнозирования.



Илл. 1.1. Процесс машинного обучения для алгоритма, отличающего кошку от собаки

Процесс обучения на основе данных и последующего применения полученных знаний для обоснования будущих решений — чрезвычайно мощный инструмент. Машинное обучение быстро превращается в двигатель современной экономики, управляемой данными.

В табл. 1.1 перечислены широко распространенные техники машинного обучения с учителем и варианты их практического применения. Список далеко не исчерпывающий, так как потенциальные варианты использования могут занять несколько страниц.

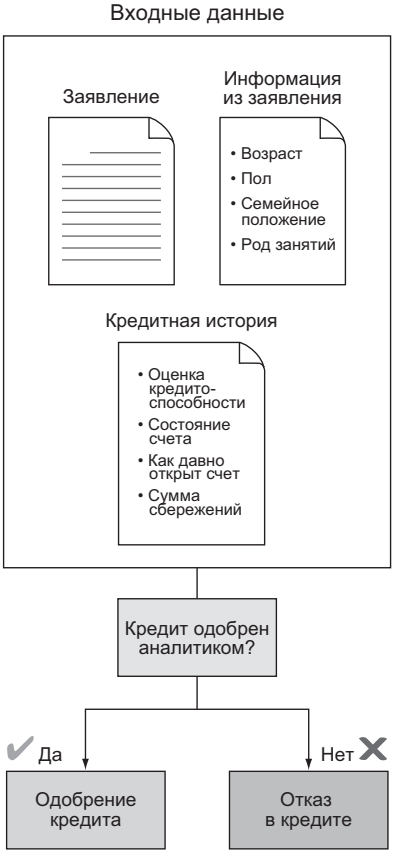
Таблица 1.1. Варианты применения машинного обучения с учителем, систематизированные по типам задач

Задача	Описание	Пример применения
Классификация	На основе данных определяется дискретный класс для каждого объекта	Фильтрация спама, анализ тональности текстов, обнаружение мошенничества, рассылка целевых рекламных объявлений, прогнозирование оттока клиентов, обработка заявок на техническую поддержку, персонализация контента, выявление производственных дефектов, сегментация потребителей, обнаружение событий, изучение геномов и эффективности лекарственных средств
Регрессия	На основе данных предсказывается фактическое значение параметра	Прогнозы на рынке ценных бумаг, прогноз спроса, прогноз цены, оптимизация аукциона реальных объявлений, управление рисками, управление активами, прогнозы погоды, спортивные предсказания
Рекомендация	Предсказывается альтернатива, которую предпочтет пользователь	Предложения продуктов, подбор персонала, конкурс Netflix Prize, онлайн-знакомства, предложение контента
Заполнение пропусков	Вывод значений отсутствующих входных данных	Неполные истории болезни, отсутствующая информация о клиентах, данные переписей

1.2. Принятие решений на основе данных

В качестве примера рассмотрим реальную экономическую задачу, решение которой упрощается с помощью машинного обучения. Мы перечислим распространенные альтернативы и покажем преимущества ML-подхода.

Представьте, что вы сотрудник организации, которая предоставляет кредиты физическим лицам для открытия малого бизнеса в неблагоприятных районах. На начальном этапе вы получали несколько заявлений в неделю, поэтому могли вручную прочитать каждое и выполнить все проверки для принятия решения об одобрении кредита. Этот процесс схематично представлен на илл. 1.2. Заемщиков устраивала скорость



Илл. 1.2. Схема одобрения кредита в микрофинансовой организации

вашей работы и индивидуальный подход, и они везде рекомендовали вашу фирму.

По мере роста популярности фирмы количество заявлений увеличивается, их уже сотни в неделю. Даже сверхурочная работа не позволяет справиться с таким наплывом, клиенты устают ждать и отправляются в конкурирующую фирму. Очевидно, что обрабатывать заявления вручную нерационально, к тому же вы начинаете испытывать сильный стресс из-за отставания от графика.

Как выйти из этой ситуации? Сейчас мы рассмотрим несколько способов, позволяющих ускорить процесс анализа заявлений, к которым обычно прибегают в подобных случаях.

1.2.1. Традиционные подходы

Исследуем два традиционных подхода к анализу данных при рассмотрении заявлений — анализ вручную и бизнес-правила. Внимательно изучим реализацию обеих техник и покажем, что именно мешает расширению бизнеса.

Увеличение штата

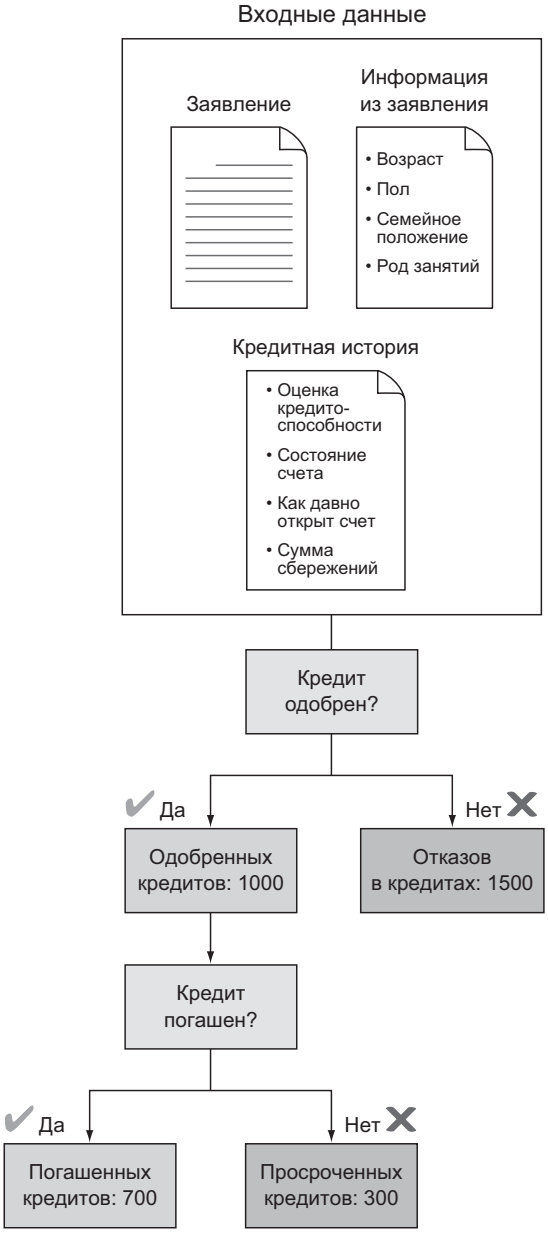
Вы нанимаете для работы с заявлениями еще одного аналитика. Тот факт, что часть прибыли придется потратить на зарплату нового сотрудника, не вызывает особого восторга, зато вдвоем можно за то же время сделать в два раза больше. В результате за неделю вы справляетесь с очередью заявок.

Пару недель ваш дуэт успевает за спросом. Но количество потенциальных клиентов продолжает расти и в следующем месяце достигает 1000 в неделю. Чтобы справиться с нагрузкой, нужны еще два аналитика. Вы делаете вывод, что в долгосрочной перспективе такая схема не работает: доход от новых заемщиков пойдет на зарплату новым сотрудникам, а не в фонд кредитования. *Увеличение штата по мере роста спроса препятствует развитию бизнеса*. Более того, сам процесс найма — длительное и дорогое развлечение, лишающее бизнес изрядной части дохода. Наконец, новый сотрудник имеет меньше опыта и обрабатывает заявки медленнее, а вы начинаете чувствовать стресс из-за необходимости управлять рабочей группой.

Кроме такого явно негативного фактора, как рост издержек, вы столкнетесь с тем, что люди добавляют в процесс принятия решений собственные сознательные и подсознательные представления. Для обеспечения согласованности потребуется детально проработать всю процедуру одобрения, а также придумать обширную программу обучения для новых аналитиков, но это еще больше увеличит издержки и вовсе не факт, что решит проблему.

Использование бизнес-правил

Представим, что из 1000 кредитов с просроченной датой погашения только 70% погашены вовремя. Эта ситуация представлена на илл. 1.3.



Илл. 1.3. За несколько месяцев работы из 2500 заявлений на кредит одобрена 1000. Из них 700 заявителей погасили кредит вовремя, а остальные 300 просрочили погашение. Этот исходный набор данных крайне важен для автоматизации процесса одобрения кредита

Теперь можно приступить к поиску связей между данными заявителя и количеством погашенных кредитов. В частности, ищется набор правил фильтрации, которые на выходе дают подмножество «хороших» кредитов, оплаченных преимущественно вовремя. Вручную проанализировав сотни заявок, вы приобретете огромный опыт, позволяющий отличать хорошую заявку от плохой.[1] После проверки накопившихся данных о погашениях кредитов вы обнаружите в процедуре проверки кредитной надежности определенные тенденции:[2]

- большинство заемщиков с лимитом кредитования более $7500 не выполняли своих обязательств по кредиту;
- большинство заемщиков, не имеющих контокоррентного счета, погашали кредит в срок.

Теперь можно спроектировать механизм фильтрации, который уменьшит количество заявлений, вручную проверяемых по двум вышеуказанным критериям.

Первый фильтр будет автоматически отказывать всем заемщикам с лимитом кредитования более $7500. Ведь накопленные данные показали, что 44 из 86 заемщиков, взявших кредит, превышающий $7500, просрочили дату его погашения. Примерно 51% из просивших максимально возможный кредит не выполнили своих обязательств по сравнению с остальными 28%. Этот фильтр кажется хорошим способом отсечения заемщиков с высоким риском невозврата. Но следует учесть, что о столь крупном кредите шла речь только в 8,6% (86 из 1000) одобренных заявок. Это значит, что более 90% анкет вам все равно придется обрабатывать вручную. То есть требуется дополнительная фильтрация.

Второй фильтр автоматически принимает любого заявителя, у которого отсутствует контокоррентный счет. Это кажется отличным решением, так как вовремя погасило свой кредит 348 из 394 (88%) заемщиков, не имеющих контокоррентного счета. Добавление этого фильтра увеличивает количество автоматически принимаемых или отклоняемых заявлений

---

[1] Для определения атрибутов исходных данных, сильнее всего связанных с итоговым событием погашения кредита, можно также использовать методы статистической корреляции.

[2] В этом примере использовался набор данных *German Credit Data*. Он доступен на странице *http://mng.bz/95r4*.