*Моей жене*

# Содержание

**От издательства** ..... 10

**О книге** ..... 11

**Глава 1. Основы работы с данными** ..... 18

1.1. Игровой автомат ..... 19
1.2. Данные с точки зрения теории вероятностей ..... 21
События и исходы ..... 21
Вероятности ..... 23
Зависимость и независимость ..... 24
Условные вероятности ..... 29
Случайные величины ..... 32
Типы случайных величин ..... 37
1.3. Закон больших чисел ..... 45
Выборочное среднее и среднее значение совокупности ..... 46
Предсказание ..... 47
Оценка параметров ..... 49
Ограничения ..... 51
1.4. Заключение и выводы ..... 52
Список литературы ..... 52

**Глава 2. Разведочный анализ данных** ..... 54

2.1. Фотонное излучение Солнца ..... 55
2.2. Сводные статистические характеристики ..... 56
2.3. Столбчатые диаграммы и гистограммы ..... 66
Столбчатые диаграммы ..... 66
Гистограммы ..... 68
2.4. Квантильный график ..... 70
Эффект упорядочивания наблюдений ..... 71
Распространенные шаблоны в квантильных графиках ..... 75
2.5. Заключение и выводы ..... 88
Список литературы ..... 89

**Глава 3. Обучение без учителя** ..... 90

3.1. Деревья и кустарники ..... 91
3.2. Ограничения базового исследования данных ..... 92
3.3. Анализ главных компонент ..... 96
Предпосылка: стандартизация данных ..... 96

Направление первой главной компоненты ... 102
Первая главная компонента ... 108
Остальные направления и главные компоненты ... 111
Наш набор данных с растениями ... 115
3.4. Кластеризация методом k-средних ... 118
Кластеризация наблюдений ... 118
Формализация кластеризации наблюдений методом k-средних ... 121
Применение к нашему примеру с растениями ... 127
Вычислительные сложности ... 131
3.5. Заключение и выводы ... 133
Список литературы ... 134

**Глава 4. Линейная регрессия** ... 135
4.1. Мощность двигателя и скорость автомобиля ... 136
4.2. Простая линейная регрессия ... 138
Детерминированные модели ... 138
Модели простой линейной регрессии ... 140
Оценивание параметров ... 141
Предсказание ... 145
Преобразование данных ... 146
4.3. Множественная линейная регрессия ... 148
4.4. Оценка качества моделей ... 151
RMSE ... 151
Коэффициент детерминации ... 153
Скорректированный $R^2$ и отложенная выборка ... 156
4.5. Заключение и выводы ... 160
Список литературы ... 162

**Глава 5. Логистическая регрессия** ... 163
5.1. Катастрофа шаттла «Челленджер» ... 164
5.2. Простая логистическая регрессия ... 166
Проблема использования линейной регрессии при наличии бинарных выходов ... 167
Более подходящая модель ... 169
5.3. Оценивание параметров и предсказание ... 172
Численная нестабильность метода наименьших квадратов ... 172
Логистический способ оценивания параметров ... 175
Предсказание ... 178
5.4. Множественная логистическая регрессия ... 179
5.5. Заключение и выводы ... 182
Список литературы ... 183

**Глава 6. Регуляризация** ... 185
6.1. Отслеживание активности ... 186
6.2. Переобучение ... 190

Пример переобучения....190
Наш пример с отслеживанием активности....194
6.3. $\ell_1$- и $\ell_2$-регуляризация....197
$\ell_2$-регуляризация уменьшает оценки параметров....197
$\ell_1$-регуляризация порождает разреженные модели....200
Регуляризация и переобучение на нашем примере отслеживания активности пользователя смартфона....203
Стандартизация данных способна облегчить регуляризацию и расчеты....205
6.4. Калибровка коэффициента регуляризации....208
Коэффициент регуляризации определяет ее интенсивность....208
Калибровка коэффициента регуляризации при помощи кросс-валидации....210
Польза от применения кросс-валидации в примере с отслеживанием активности....215
6.5. Заключение и выводы....216
Список литературы....216

**Глава 7. Метод опорных векторов**....218

7.1. Футбол....219
7.2. Вступление: гиперплоскости....222
Гиперплоскости – это аффинные ($d$ – 1)-мерные подмножества пространства признаков....224
Гиперплоскости разделяют пространство признаков надвое....227
Гиперплоскости представляют правила классификации....231
7.3. Классификаторы на основе опорных векторов и метод опорных векторов....232
Пример: овцы и волки....232
Выбор гиперплоскости при помощи классификатора на основе опорных векторов....234
Продолжение примера о футболе....253
Машины опорных векторов как способ обобщения классификатора опорных векторов....255
7.4. Расширения: веса и добавление классов....267
Веса классов....267
Многоклассовая классификация....271
7.5. Заключение и выводы....278
Список литературы....279

**Глава 8. Глубокое обучение**....281

8.1. FashionMNIST....282
8.2. Перцептрон....285
Нейроны перцептрона....287
Алгоритм перцептрона....290
Сравнение перцептрона с классификатором на основе опорных векторов....294

8.3. От нейронов к глубокому обучению ..... 295
Неглубокая нейронная сеть ..... 295
Оценка параметров ..... 307
Сравнение с классификатором на основе опорных векторов ..... 314
Глубокое обучение ..... 315
8.4. Заключение и выводы ..... 321
Список литературы ..... 322

**Предметный указатель** ..... 324

# От издательства

## *Отзывы и пожелания*

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте *www.dmkpress.com*, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу *dmkpress@gmail.com*; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу *http://dmkpress.com/authors/publish_book/* или напишите в издательство по адресу *dmkpress@gmail.com*.

## *Список опечаток*

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу *dmkpress@gmail.com*. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

## *Нарушение авторских прав*

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты *dmkpress@gmail.com*.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

# О книге

## О чем эта книга?

Наука о данных представляет собой вид искусства для превращения сырых необработанных данных в ценную и значимую информацию. Внедрение науки о данных на основе принципов статистики и называется *статистическим обучением* (statistical learning). Статистическое обучение устанавливает систематические правила для данных и описывает отклонения от этих правил. В этой книге вы познакомитесь с основами статистического обучения и его наиболее распространенными методами. Книга условно состоит из трех частей.

Первые три главы посвящены **данным** как основе статистического обучения. Нашим математическим инструментарием для работы с данными является *теория вероятностей* (probability theory), позволяющая моделировать изменения в данных в соответствии с нашими интуитивными представлениями о шансах. Затем данные описываются с помощью базовых сводных статистик (среднее, количество наблюдений и т. д.), визуализаций, таких как гистограмма или квантильный график, и методов обучения без учителя (анализ главных компонент, метод $k$-средних и пр.).

- Глава 1. Основы работы с данными.
- Глава 2. Разведочный анализ данных.
- Глава 3. Обучение без учителя.

**Инференциальный анализ данных**, лежащий в основе второй части книги, позволяет сделать выводы относительно процесса генерирования данных. Такой анализ подразумевает использование статистических моделей, таких как линейная и логистическая регрессия.

- Глава 4. Линейная регрессия.
- Глава 5. Логистическая регрессия.
- Глава 6. Регуляризация.

Сделать предсказания относительно новых данных позволяют методы машинного обучения. Эти предсказания могут включать в себя статистические модели, но сами модели здесь – не главное.

- Глава 7. Метод опорных векторов.
- Глава 8. Глубокое обучение.

Мы увидим, как все три части взаимопроникают друг в друга. К примеру, с помощью методов обучения без учителя можно получить признаки для

инференциального анализа данных, линейная или логистическая регрессия может помочь в прогнозировании новых данных, а регуляризация может использоваться как в методах обучения без учителя, так и в инференциальном анализе данных.

# О переводчике


**Александр Гинько**, обладающий богатым опытом работы в сфере ИТ и более 15 лет посвятивший переводам книг и статей на самые разные темы, в последние годы специализируется на переводе книг в области бизнес-аналитики и программирования для издательства «ДМК Пресс» по направлениям Python, SQL, R, Power BI, DAX, статистика, машинное обучение, нейронные сети, Excel, Power Query, Tableau, ... На данный момент в активе Александра уже порядка 30 книг, включая одну авторскую, и он продолжает плодотворно работать над переводом новых книг и написанием собственных.

Возможно, вам также будут интересны книги «Введение в статистическое обучение с примерами на Python» (*https://dmkpress.com/catalog/computer/statistics/978-5-93700-217-4*) и «Машинное обучение сквозь призму Excel. Примеры и упражнения» (*https://dmkpress.com/catalog/computer/data/978-5-93700-238-9*) в переводе Александра.

Помимо перевода книг, Александр ведет свой канал в Telegram (*https://t.me/alexanderginko_books*), на котором вы можете из первых уст получить ответы на все интересующие вас вопросы об уже переведенных книгах, находящихся в работе и запланированных на будущее. Также на канале можно найти промокоды на все книги Александра для покупки книг на сайте издательства «ДМК Пресс» с большими скидками. Купить книги Александра и следить за переводом новых книг в режиме реального времени можно и с помощью его бота в Telegram по адресу *https://t.me/alexanderginko_books_bot*.

# Для кого эта книга?

Эта книга предназначается всем, кто хочет понимать и уметь применять на практике основные концепции и методы статистического обучения. Она может подойти студентам 3–4 курсов института или учащимся 1–2 курсов магистратуры по специальностям, тесно связанным с данными, таким как

компьютерные науки, биология, психология, бизнес или инженерия. Также она может помочь выпускникам вузов при прохождении собеседований. Для описания идей статистического обучения недостаточно обычных слов из нашего обихода. В добавление к ним понадобится иллюстрация при помощи кода и объяснение на строгом языке математических формул. В идеале читатель этой книги должен обладать минимальным опытом программирования и быть знакомым на базовом уровне с математическими формулами для работы с множествами, случайными величинами и функциями. Но не беспокойтесь, это не книга по математике, и мы постараемся использовать как можно меньше математических выкладок. В частности, мы будем избегать упоминания сложных формулировок и вычислений, относящихся, к примеру, к векторному исчислению. В целом можно сказать, что книга, которую вы держите в руках, может идеально подойти как для самообразования, так и в качестве основы для образовательных курсов.

# Что в этой книге особенного?

В статистическом обучении теория и практика всегда идут рука об руку. Однако в большинстве обучающих материалов по данной дисциплине это не учитывается. В них либо делается упор только на реализацию (обычно это касается онлайн-курсов), что ведет к потере общего контекста и глубины постижения материала, либо концепция и применение явным образом разделяются (характерно для учебников, написанных в стиле «теория – упражнения – лабораторные работы»), в результате чего создается искусственный разрыв между этими важными аспектами. В этой книге мы постарались объединить вместе теорию, упражнения и реализацию изученных методов на практике, что поможет выдержать глубину погружения в материал и вместе с тем сохранить практическую направленность книги.

Кроме того, эта книга обладает следующими полезными свойствами:

- тщательный отбор тем, позволяющий ускорить процесс обучения;
- постановка вопроса в начале каждой главы, помогающая читателю следить за развитием мысли и поиском ответа;
- скрупулезное изложение тем, основанное при этом на элементарной математике;
- более двух сотен упражнений, которые помогут лучше понять излагаемый материал;
- полезный раздел в конце каждой главы с подведением итогов описанных концепций и закреплением их на практике;
- многочисленные рекомендации по поиску дополнительной информации для будущего изучения.

# Почему Python?

В науке о данных используется великое множество языков программирования, включая Python, Julia, Java, R и MATLAB. У каждого из них есть свои сильные и слабые стороны, а чтобы стать настоящим специалистом по анализу данных, вам наверняка понадобится освоить более одного языка. Но для начинающих наиболее подходящим языком является именно Python в силу простоты изучения, гибкости применения и наличия большого количества библиотек и обучающих материалов. Все это делает язык Python наиболее распространенным в области науки о данных, да и не только в ней.

# Как использовать эту книгу?

В тексте книги чередуются изложения концепций, код на языке Python и упражнения для самостоятельного выполнения. При чтении книги мы рекомендуем следовать плану, т. е. изучать концепции и реализовывать код на практике, а также выполнять упражнения по мере их появления. При использовании электронного варианта книги вы, конечно, можете просто копировать приведенный код, но вы, наверное, знаете, что наибольшей эффективности в процессе обучения можно добиться, только набирая весь код самостоятельно.

В каждой главе мы будем работать с одним из наборов данных, которые можно найти в сопроводительных материалах к книге на сайте издательства или загрузить по адресу *https://johannesledererer.com/statisticallearning*.

Упражнения, приведенные прямо в тексте книги, помогут вам глубже усвоить материал и посмотреть на описываемые концепции под другим углом. Упражнения будут помечаться изображениями баскетбольного мяча: 🏀. Чем больше мячей, тем сложнее упражнение. Прилагайте усилия для решения всех упражнений, встречаемых на пути, но не стоит расстраиваться, если какие-то из них вам окажутся не по зубам. Важно, чтобы упражнения заставили вас задуматься о предмете обсуждения.

Во фрагментах кода на языке Python мы будем использовать необходимые инструменты для решения стоящих перед нами задач. Также мы будем отдельно помечать значком 🌀 нестандартные приемы, которые вы можете использовать в Python. Но это не учебник по Python, так что если вам потребуется дополнительная информация по командам, функциям и методам языка, вам придется обратиться за подробностями к документации.

Также в книге вы найдете отсылки на другую полезную литературу. Уровень сложности этих материалов будет варьироваться в зависимости от количества значков 🎓: один значок будет означать простой для освоения материал, два значка – более специализированные статьи с точки зрения математических выкладок, а три значка – материалы для хорошо подкованных читателей, такие как научные и исследовательские работы.

В конце каждой главы вас будет ждать раздел с подведением итогов, а важные выводы мы будем помечать значком 💡. Наблюдения, на первый взгляд противоречащие здравому смыслу (разрушители мифов), мы будем помечать значком ✵.

Образовательный курс, в основе которого лежит материал из этой книги, обычно завершается проектом: студенты выбирают набор данных для анализа, исследуют его при помощи методов, описанных в книге, и представляют результаты в виде короткого видео. Это бывает очень полезно как для студентов, так и для преподавателей. Вы можете разрабатывать подобные проекты самостоятельно и показывать коллегам, друзьям и, самое главное, самому себе, чтобы убедиться в том, что вы научились чему-то полезному!

# Нужно ли читать главы книги в определенном порядке?

Нет. Все восемь глав в этой книге по большей части самодостаточны, а значит, вы можете читать их в любом порядке. Но у меня лично есть две рекомендации. Первая связана с тем, что во всех главах книги мы будем обращаться к базовым концепциям теории вероятностей, хотя вы можете попытаться понять материалы из них и на более интуитивном уровне. Таким образом, вам будет лучше начать чтение с первой главы, посвященной основам работы с данными, если вы хотите постичь математический фундамент статистического обучения. В то же время вы можете пропустить эту главу, если вашей целью является освоение методов и их реализаций. Кроме того, в трех главах второй части книги упоминаются одни и те же концепции, так что их я рекомендую прочитать в порядке, представленном в книге.

# Какие модули Python понадобятся?

Для написания всех фрагментов кода из этой книги вам достаточно будет установить следующие версии девяти пакетов:

```
Python 3.10.6
numpy 1.23.5
pandas 1.5.2
matplotlib 3.6.2
scipy 1.9.3
seaborn 0.12.0
sklearn 1.1.2
plotly 5.10.0
torch 1.13.0
torchvision 0.14.0
```

Также я рекомендую использовать среду разработки JupyterLab, особенно если вы работаете на Windows. Код был написан и протестирован в JupyterLab версии 3.4.7.

# Математические обозначения, используемые в книге

При написании книги я постарался по максимуму, насколько это возможно, ограничить присутствие математических формул. В то же время в статистике без формул не обойтись, так что мы должны определиться с некоторыми основными способами обозначения.

## Числа и параметры

*Статистические параметры* (statistical parameter) мы будем обозначать греческой буквой $\beta$, а соответствующие им *оценки* (estimator) – $\hat{\beta}$. Остальные числа мы будем обозначать строчными буквами с обычным шрифтом: $a$. Символ $\approx$ обозначает приблизительное равенство. Например, запись $\pi \approx 3.141$ означает, что константа $\pi = 3.14159265359\ldots$ приблизительно равна 3.141. Числа, взятые по модулю, мы будем заключать в символы вертикальной черты. Пример: $|1| = |-1| = 1$.

## Множества и их элементы

*Множества* (set) мы будем обозначать заглавными буквами каллиграфического шрифта: $\mathcal{A}$. В качестве исключения множество вещественных чисел обозначается при помощи буквы $\mathbb{R}$. Символом $\in$ обозначается вхождение элемента в множество. К примеру, запись $2 \in \{1, 2, 3\}$ означает, что число 2 является элементом множества $\{1, 2, 3\}$. С помощью символа # обозначается кардинальность (количество элементов), или мощность, множества. Пример: $\#\{0, 1, 2\} = 3$. Обозначения $\min \mathcal{A}$ и $\max \mathcal{A}$ служат для указания на минимальный и максимальный элементы множества соответственно. Примеры: $\min \{0, 1\} = 0$, $\max \{0, 1\} = 1$.

## Определения и эквивалентность

Символом := обозначается фраза «определен как». Например, запись $a := 1$ означает, что мы присваиваем $a$ значение 1. Символом $\Leftrightarrow$ обозначается эквивалентность значений. Пример: $a = 1 \Leftrightarrow 2a = 2$ означает, что выражения $a = 1$ и $2a = 2$ просто переформулируют друг друга.

## Функции

*Вероятностная мера* (probability measure), *кумулятивная функция распределения* (cumulative distribution function) и *ожидание* (expectation) обозначаются

$\mathbb{P}$, $\mathbb{F}$ и $\mathbb{E}$ соответственно. Случайные величины мы будем обозначать заглавными буквами с обычным шрифтом, например $Y$. Натуральный логарифм записывается как log и означает логарифм по основанию числа Эйлера, или $e$, где $e \approx 2.718$. Пример: $\log[e] = 1$. Все остальные веществennознозначные функции обозначаются строчными буквами готического шрифта: $\mathfrak{f}$.

### *Минимум и максимум функций*

Минимум функции $\mathfrak{f}$ из множества $\mathcal{X}$ обозначается как $\min_{x \in \mathcal{X}} \mathfrak{f}[x]$. Минимум может представлять собой как уникальное число, так и множество чисел, а также пустое множество (при отсутствии минимума): $\min_{x \in \mathbb{R}} x^2 = 0$, $\min_{x \in \mathbb{R}} (x + 1/2)^2/(x - 1/2)^2/(3/4)^2 = 0$, а $\min_{x \in \mathbb{R}} e^{-x-1}$ – пустое множество. Аргументы, минимизирующие функцию $\mathfrak{f}$ на множестве $\mathcal{X}$, обозначаются как $\operatorname{argmin}_{x \in \mathcal{X}} \mathfrak{f}[x]$. Это может быть один аргумент, множество аргументов или их отсутствие: $\operatorname{argmin}_{x \in \mathbb{R}} x^2 = 0$, $\operatorname{argmin}_{x \in \mathbb{R}} (x + 1/2)^2/(x - 1/2)^2/(3/4)^2 = \{-1/2, 1/2\}$, а $\operatorname{argmin}_{x \in \mathbb{R}} e^{-x-1}$ – пустое множество. Так же точно максимум функции $\mathfrak{f}$ из множества $\mathcal{X}$ обозначается как $\max_{x \in \mathcal{X}} \mathfrak{f}[x]$, а аргументы, максимизирующие функцию $\mathfrak{f}$ на множестве $\mathcal{X}$, обозначаются как $\operatorname{argmax}_{x \in \mathcal{X}} \mathfrak{f}[x]$. На рис. 0.1 показаны три примера функции (жирные синие кривые), а также их минимумы (черные горизонтальные линии) и соответствующие минимизирующие аргументы (красные вертикальные линии).

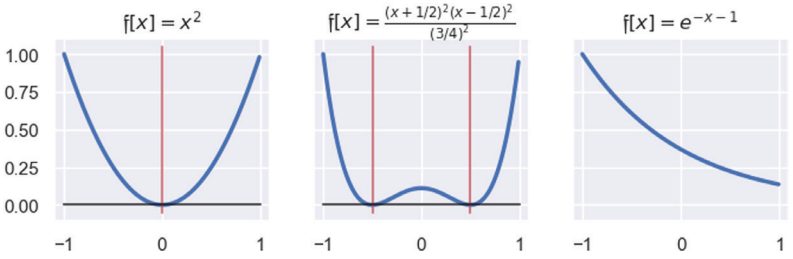


**Рис. 0.1** ❖ Функции, их минимумы и минимизирующие аргументы

# Благодарности

Я выражаю свою искреннюю благодарность Милене Брейг (Milena Braig), Дюку Лонгу Чу (Duc Long Chu), Пеге Голестане (Pegah Golestaneh), Ших-Тингу Хуангу (Shih-Ting Huang), Ларе Мейн (Lara Meyn), Али Мохаддесу (Ali Mohaddes) и Махсе Тахери (Mahsa Taheri) за их ценные советы и предложения в процессе написания книги. Также я очень признателен Веронике Ростек (Veronika Rosteck) из издательства Springer за прекрасное руководство проектом написания этой книги.

# Глава 1

# Основы работы с данными

**Выгодно ли играть в азартные игры[1]?**

Ответ на этот вопрос вы найдете в этой главе.

Исходным материалом для статистического обучения являются *данные*. При этом разные приложения, задействующие инструменты статистического обучения, могут использовать разные данные: от котировок акций на бирже до замеров амплитуды пения птиц и уровней нейронной активности. Несмотря на кардинальные различия в исходных данных, с помощью

[1] Фотография сделана Йоханнесом Ледерером.

методов статистического обучения можно досконально проанализировать их и выявить, например, повышенную волатильность ценных бумаг после минувшего финансового кризиса или определить степень влияния пандемии коронавируса на пение птиц, хотя между этими двумя областями, как вы понимаете, настоящая пропасть[2]. В основе такой универсальности статистического обучения лежит тот факт, что его методы формализуют данные в абстрактной математической оболочке, именуемой теорией вероятностей. В этой части книги мы познакомимся с данной оболочкой и научимся использовать ее для описания исходных данных.

В первой главе мы узнаем, как данные трансформируются в случайные величины, представляющие ключевую концепцию в теории вероятностей, и обсудим важные свойства этих случайных величин. Вы научитесь:

- **формулировать** данные на языке теории вероятностей;
- **различать** разные типы случайных величин;
- **извлекать** информацию из данных.

# 1.1. Игровой автомат

Ни для кого не секрет, как работают игровые автоматы, которые можно встретить в казино и других специализированных заведениях. Вы бросаете монетку и проверяете свое везение. Если вам повезет, автомат расщедрится на бесчисленное количество монет, сопровождая их выдачу громкими звуками и миганием лампочек. Вопрос прост: у кого больше шансов обогатиться: у вас или у владельца автомата?

В файле `Gambling.csv`, который вы можете найти в сопроводительных материалах к книге, мы привели (вымышленные) данные об исходах игры на таком автомате. Каждая игра стоит нашему игроку 0.25$, а автомат в зависимости от везения игрока и каких-то своих алгоритмов может либо молча проглотить его монету с приглашением попытать удачу снова, либо выдать 20 монет в качестве приза (т. е. исхода может быть два: 0$ (для игрока это −0.25$) или 5$ (для игрока это +4.75$)). В первом столбце в файле содержится исход игры, если допустить, что игрок играл ровно один раз в день в течение года. В случае выигрыша мы увидим строку `data scientist wins`, а в случае проигрыша – `data scientist loses`. Во втором столбце – чистая прибыль игрового автомата по всем играм за каждый день. Фрагмент файла показан ниже:

```
# Individual,Machine
data scientist loses,46.750000
data scientist loses,37.500000
data scientist loses,55.250000
```

[2] Вы сможете вернуться к этим вопросам после прочтения следующей главы книги. См. 🎓 Derryberry, E., Phillips, J., Derryberry, G., Blum, M. and Luther, D., 2020. Singing in *a* silent spring: birds respond to *a* half-century soundscape reversion during the COVID-19 shutdown. Science, 370 (6516), pp. 575–579.

```
data scientist wins,32.750000
…
data scientist loses,4.250000
data scientist loses,56.000000
data scientist loses,29.750000
data scientist loses,30.250000
data scientist loses,37.000000

import numpy as np
records = np.genfromtxt("Gambling.csv", delimiter=",", dtype=None, encoding=None)
print(records[:5,] ) # пример из нескольких записей
```

Вывод:

```
[('data scientist loses', 46.75) ('data scientist loses', 37.5 )
('data scientist loses', 55.25) ('data scientist wins', 32.75)
('data scientist loses', 14. )]
```

Аргументы `dtype=None` и `encoding=None` позволяют строкам из первого столбца загружаться корректно.

Как видим, в первые пять дней игрок терпел неудачу четыре раза из пяти, а в один из дней сорвал джекпот. В то же время чистая прибыль автомата в эти дни варьировалась от 14.00$ до 55.25$.

По этому небольшому фрагменту данных можно сделать вывод, что и наш игрок, и игровой автомат неплохо заработали за первые пять дней общения: чистая прибыль игрока составила –0.25$ – 0.25$ – 0.25$ + 4.75$ – 0.25$ = 3.75$, а чистая прибыль владельца автомата – 46.75$ + 37.50$ + 55.25$+32.75$+14.00$ = 186.25$. Но за весь год результаты выглядят иначе:

```
ds_records = [i for i, j in records]
ds_net_earning = (4.75 * ds_records.count("data scientist wins")
                  - 0.25 * ds_records.count("data scientist loses"))
print(f"Игрок проиграл всего: {-ds_net_earning}$.")

machine_records = [j for i, j in records]
print(f"Автомат выиграл всего: {np.sum(machine_records)}$.")
```

Вывод:

```
Игрок проиграл всего: 31.25$.
Автомат выиграл всего: 13627.75$.
```

Выражения `[i for i, j in records]` и `[j for i, j in records]` являются примером *генератора списков* (list comprehension)[3]. Более многословный аналог этого кода представлен ниже:

[3] Подробнее о генераторах списков и прочих генерирующих выражениях можно почитать в главе 9 издания Phillips, D., 2018. Python 3 object-oriented programming: build robust and maintainable software with object-oriented design patterns in Python 3.8. Third edition. Packt.