

УДК 004.4  
ББК 32.972  
К62

**Кольер Р., Монтонен К., Азарми Б.**

К62 Машинное обучение в Elastic Stack / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2022. – 380 с.: ил.

**ISBN 978-5-93700-107-8**

В книге подробно рассматривается работа с Elastic Stack – обширной экосистемой компонентов, которые служат для сбора, поиска и обработки данных. Вы ознакомитесь с общими принципами машинного обучения, узнаете о методах автоматического обнаружения аномалий, проверке целостности и анализа данных из разрозненных источников, научитесь истолковывать результаты обнаружения и прогнозирования аномалий и использовать их в своих целях, а также выполнять анализ временных рядов для различных типов данных.

Издание адресовано специалистам, которые работают с данными и хотят интегрировать машинное обучение с эффективными приложениями для мониторинга, обеспечения безопасности и аналитики в области данных.

УДК 004.4  
ББК 32.972

First published in the English language under the title 'Machine Learning with the Elastic Stack. Second Edition (978-1-80107-003-4). The Russian-Language 1st edition Copyright © 2021 by DMK Press Publishing under license by No Starch Press Inc. All rights reserved.

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.

ISBN 978-1-80107-003-4 (англ.)  
ISBN 978-5-93700-107-8 (рус.)

© Packt Publishing, 2021  
© Перевод, оформление, издание,  
ДМК Пресс, 2022

# Содержание

От издательства .....	12
Об авторах .....	13
О рецензентах .....	14
Предисловие .....	15

<b>Часть I. ЗНАКОМСТВО С МАШИНЫМ ОБУЧЕНИЕМ И ELASTIC STACK .....</b>	<b>18</b>
--	-----------

<b>Глава 1. Машинное обучение в информационных технологиях.....</b>	<b>19</b>
Преодоление исторических вызовов в IT .....	19
Что нам делать с потоком данных? .....	20
Причины появления автоматического обнаружения аномалий.....	21
Машинное обучение без учителя и с учителем.....	23
Использование машинного обучения без учителя для обнаружения аномалий .....	24
Что такое необычность?.....	24
Изучение того, что является нормой .....	26
Вероятностные модели .....	26
Обучение моделей .....	27
Выявление и устранение тенденций .....	30
Оценка степени необычности .....	31
Роль времени .....	32
Применение машинного обучения с учителем в аналитике фреймов данных .....	33
Процесс обучения с учителем.....	33
Заключение .....	35

<b>Глава 2. Подготовка и использование Elastic ML.....</b>	<b>36</b>
Технические требования.....	36
Включение функций Elastic ML.....	36
Включение машинного обучения в собственном кластере .....	37
Включение машинного обучения в облаке – Elasticsearch Service .....	39
Обзор операционализации Elastic ML .....	46
Узлы ML .....	46
Задания.....	47
Сегментирование данных в анализе временных рядов .....	48

Загрузка данных в Elastic ML .....	49
Служебные хранилища .....	51
.ml-config.....	51
.ml-state-* .....	51
.ml-notifications-* .....	52
.ml-annotations-* .....	52
.ml-stats-* .....	52
.ml-anomalies-* .....	52
Оркестровка обнаружения аномалий.....	52
Снимки модели обнаружения аномалий.....	53
Заключение .....	54

## **Часть II. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ – ОБНАРУЖЕНИЕ И ПРОГНОЗИРОВАНИЕ АНОМАЛИЙ .....**

### **Глава 3. Обнаружение аномалий.....**

Технические требования.....	56
Типы заданий Elastic ML.....	56
Устройство детектора .....	58
Функция .....	59
Поле.....	59
Поле partition .....	59
Поле by .....	60
Поле over .....	60
Формула детектора.....	60
Обнаружение изменений частотности событий.....	61
Подробнее о функциях count .....	61
Другие функции подсчета .....	73
Ненулевой подсчет .....	73
Раздельный подсчет .....	74
Обнаружение изменений значений показателей.....	75
Метрические функции .....	76
min, max, mean, median и metric.....	76
varp.....	76
sum и non-null sum .....	76
Обзор расширенных функций детектора.....	77
Функция rare.....	78
Функция freq_rare .....	79
Функция info_content.....	79
Функции геолокации.....	79
Функции времени.....	80
Разделение анализа по категориальным признакам.....	80
Настройка поля разделения .....	80
Разница между разделением с использованием partition и by_field .....	82
Является ли двойное разделение пределом возможного?.....	83
Обзор временного и популяционного анализов.....	84

Категоризация в анализе неструктурированных сообщений.....	86
Типы сообщений, подходящие для категоризации.....	88
Предварительная категоризация .....	88
Анализ категорий .....	89
Пример задания по категоризации .....	90
Когда не следует использовать категоризацию .....	94
Управление Elastic ML через API .....	95
Заключение .....	97

## **Глава 4. Прогнозирование..... 98**

Технические требования.....	98
Ключевое различие между предсказаниями и пророчествами.....	98
Для чего применяется прогнозирование? .....	100
Как работает прогнозирование?.....	100
Прогнозирование одиночного временного ряда.....	103
Просмотр результатов прогнозирования.....	114
Прогнозирование нескольких временных рядов .....	119
Заключение .....	122

## **Глава 5. Интерпретация результатов..... 123**

Технические требования.....	123
Просмотр хранилища результатов Elastic ML.....	123
Оценка аномалий.....	128
Оценка на уровне сегмента.....	129
Нормализация .....	131
Оценка на уровне фактора влияния .....	131
Факторы влияния.....	133
Оценка на уровне записи .....	135
Описание схемы хранилища результатов.....	136
Результаты на уровне сегмента .....	137
Результаты на уровне записи.....	140
Результаты на уровне факторов влияния .....	143
Аномалии в нескольких сегментах .....	145
Пример аномалии в нескольких сегментах.....	145
Оценка аномалии в нескольких сегментах.....	146
Результаты прогноза .....	148
Запрос результатов прогноза.....	149
API результатов Elastic ML.....	151
Конечные точки API результатов .....	152
API обобщения сегментов .....	152
API категорий.....	153
Пользовательские панели мониторинга и рабочие панели Canvas .....	155
Панель инструментов встраивания .....	155
Аномалии как аннотации в TSVB.....	156
Настройка рабочих панелей Canvas.....	159
Заключение .....	162

<b>Глава 6. Создание и использование оповещений</b>	163
Технические требования	163
Определение и принцип работы оповещений	164
Аномалии не обязательно нуждаются в оповещениях	164
Точное время имеет значение	165
Создание оповещений из интерфейса машинного обучения	168
Определение заданий по обнаружению аномалий	168
Создание оповещений для пробных заданий	174
Моделирование аномального поведения в реальном времени	179
Получение и просмотр оповещений	180
Создание оповещений с помощью Watcher	183
Использование устаревшего варианта watch	183
trigger	184
input	184
condition	187
action	188
Пользовательские шаблоны watch с уникальной функциональностью	189
Связанный ввод и сценарий условий	189
Передача информации между связанными входами	190
Заключение	191
<b>Глава 7. Выявление истинных причин аномалий</b>	192
Технические требования	192
Настоящее значение термина AIOps	192
Значимость и ограничения KPI	194
Выходя за рамки KPI	197
Организация данных для анализа	198
Настраиваемые запросы для каналов данных	199
Дополнение получаемых данных	202
Использование контекстной информации	203
Аналитическое разделение	203
Статистические факторы влияния	204
Анализ первопричин аномалии	205
История проблемы	205
Корреляция и общие факторы влияния	207
Заключение	212
<b>Глава 8. Другие приложения Elastic Stack для обнаружения аномалий</b>	213
Технические требования	213
Обнаружение аномалий в Elastic APM	213
Включение обнаружения аномалий для APM	214
Просмотр результатов задания по обнаружению аномалий	219
Создание заданий машинного обучения с помощью распознавателя данных	222
Обнаружение аномалий в приложении Logs	224

Категории журналов.....	224
Журнал аномалий .....	225
Обнаружение аномалий в приложении Metrics .....	227
Обнаружение аномалий в приложении Uptime .....	230
Обнаружение аномалий в приложении Elastic Security .....	233
Готовые задания по обнаружению аномалий .....	233
Оповещения на основе заданий обнаружения аномалий.....	235
Заключение .....	237

## **Часть III. АНАЛИЗ ФРЕЙМОВ ДАННЫХ ..... 238**

### **Глава 9. Введение в анализ фреймов данных..... 239**

Технические требования.....	240
Основы преобразования данных.....	240
Чем полезны преобразования?.....	240
Анатомия преобразований .....	241
Использование преобразований для анализа заказов интернет-магазина.....	242
Более сложные конфигурации сводной таблицы и агрегирования .....	246
Различие между пакетными и непрерывными преобразованиями.....	248
Анализ данных социальных сетей с помощью непрерывных преобразований .....	250
Использование Painless для расширенных конфигураций преобразования.....	253
Знакомство с Painless .....	254
Переменные, операторы и управление выполнением .....	255
Функции.....	260
Совместное использование Python и Elasticsearch.....	263
Краткий обзор клиентов Python Elasticsearch.....	264
Зачем нам нужен Eland? .....	266
Знакомство с Eland .....	267
Заключение .....	269
Дополнительная литература .....	270

### **Глава 10. Обнаружение выбросов ..... 272**

Технические требования.....	273
Принцип работы механизма обнаружения выбросов.....	273
Обзор четырех методов обнаружения выбросов .....	274
Методы, основанные на расстоянии .....	274
Методы, основанные на плотности .....	275
Оценка влияния характеристики.....	276
Как рассчитывается оценка влияния характеристик для каждой точки?.....	277
Чем обнаружение выбросов отличается от обнаружения аномалий? .....	278
Сравнение вероятностных моделей и экземпляров .....	278
Подсчет оценок.....	279

Характеристики данных.....	279
Потоковая и пакетная обработка.....	279
Применение обнаружения выбросов на практике .....	280
Оценка качества обнаружения выбросов с помощью API Evaluate .....	285
Настройка гиперпараметров для обнаружения выбросов .....	290
Заключение .....	293
<b>Глава 11. Классификационный анализ .....</b>	<b>294</b>
Технические требования.....	295
Классификация: от данных к обученной модели.....	295
Классифицирующие модели учатся на данных .....	296
Конструирование признаков .....	298
Оценка модели .....	299
Простой пример классификации.....	300
Деревья решений с градиентным усилением.....	307
Введение в деревья решений .....	308
Градиентное усиление .....	309
Гиперпараметры .....	309
Интерпретация результатов.....	313
Вероятность класса.....	314
Оценка класса .....	314
Важность признака.....	314
Заключение .....	316
Дополнительная литература .....	317
<b>Глава 12. Регрессия.....</b>	<b>318</b>
Технические требования.....	318
Использование регрессионного анализа для прогнозирования цен на жилье.....	319
Использование деревьев решений в регрессионном анализе .....	326
Заключение .....	329
Дополнительная литература .....	329
<b>Глава 13. Логический вывод моделей .....</b>	<b>330</b>
Технические требования.....	330
Поиск, импорт и экспорт обученных моделей с помощью API .....	331
Обзор API обученных моделей .....	331
Экспорт и импорт обученных моделей с помощью API и Python.....	333
Обработчики логического вывода и конвейеры данных.....	336
Обработка отсутствующих или поврежденных данных в конвейерах.....	345
Получение развернутой информации о прогнозах.....	347
Импорт внешних моделей с помощью eLand.....	348
Кратко о поддержке внешних моделей в eLand.....	349
Обучение DecisionTreeClassifier и импорт в Elasticsearch с помощью eLand.....	349
Заключение .....	353

<b>Приложение. Советы по обнаружению аномалий .....</b>	<b>354</b>
Технические требования.....	354
Роль факторов влияния в разделенных и неразделенных заданиях .....	354
Использование односторонних функций .....	361
Исключение определенных интервалов времени .....	363
Исключение наступающего (известного) интервала времени .....	364
Создание события календаря .....	364
Остановка и запуск потока данных в нужное время .....	365
Исключение интервала времени постфактум.....	366
Клонирование задания и повторный запуск исторических данных.....	366
Возврат задания к предыдущему снимку модели.....	366
Использование настраиваемых правил и фильтров .....	368
Создание собственных правил .....	369
Использование настраиваемых правил для оповещения «сверху вниз».....	370
Соображения относительно пропускной способности заданий.....	371
О вреде излишней сложности сценариев.....	372
Обнаружение аномалий в вычисляемых полях .....	373
Заключение .....	376
 <b>Предметный указатель.....</b>	 <b>377</b>