

# 1. Введение

→ В предыдущих модулях мы изучили темы визуализации, работы с признаками, повторили принципы очистки данных и работу с функциями. Возникает вопрос: в каких случаях эти навыки применяет специалист по машинному обучению в своей работе? Когда мы будем обучать модели?

**EDA (Exploratory Data Analysis)** — разведывательный анализ данных. Этот этап дата-сайентисты проводят перед построением самой модели. Цель этого этапа — понять, что нам могут дать данные, и как признаки могут быть взаимосвязаны между собой. Понимание изначальных признаков позволит создать новые, более сильные признаки и повысить качество модели.

Качественная модель способна выдавать правильные результаты на любых новых данных, которые никогда не участвовали в обучении.

Изучив разведывательный анализ данных, мы будем готовы создать первую модель! *EDA* помогает искать смысл и закономерности в данных, но для этого нам необходимы все навыки, освоенные в модулях раньше!

*EDA* — это совокупность разнообразных навыков и умений, инструментов для работы с данными.

## В ЭТОМ МОДУЛЕ ВЫ УЗНАЕТЕ:

☒ какую роль играет анализ данных в машинном обучении;

## ЦЕЛИ МОДУЛЯ:

☒ Узнать, какую роль *EDA* играет в машинном обучении.

☒ какие библиотеки

*Python* помогут вам в освоении разведывательного анализа данных;

☒ поработаете с

данными и узнаете, какие типы признаков существуют;

☒ проверите знания из

прошлых модулей и примените их на практике;

☒ узнаете, как можно

сделать разведывательный анализ данных одной строчкой кода.

☒ Узнать, какие навыки

необходимы в освоении *EDA*.

☒ Узнать, какие типы

признаков существуют, и уметь их различать.

☒ Поработать с

библиотеками и инструментами, которые облегчат *EDA*.

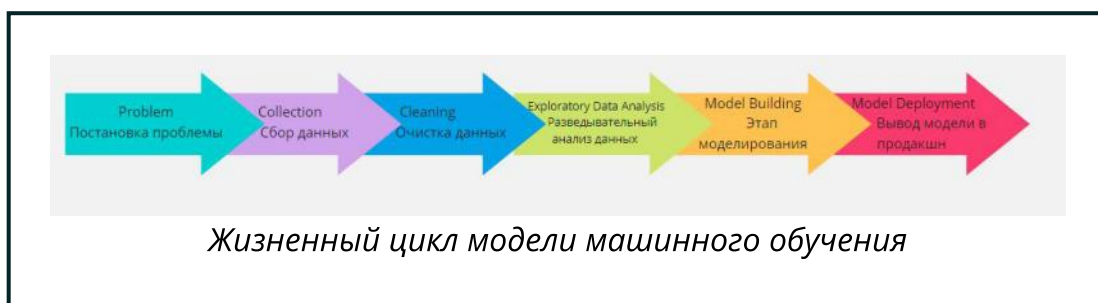
→ В следующем юните вы узнаете, какую роль разведывательный анализ данных играет в машинном обучении. А пока предлагаем закрепить пройденный материал заданиями ниже.

## 2. Жизненный цикл ML-модели и роль EDA в машинном обучении

→ В прошлом юните мы узнали, что такое *EDA*, и сформировали цели на модуль. В этом юните мы рассмотрим полный жизненный цикл модели машинного обучения.

Тот, кто занимается машинным обучением, обычно проходит следующие этапы: формулировка бизнес-проблемы, сбор данных и их очистка, разведывательный анализ данных, разработка и построение модели, внедрение модели в продакшен.

Всё это называется **жизненным циклом модели машинного обучения**.



Мы можем выделить несколько этапов жизненного цикла модели машинного обучения.

### 0. ПОСТАНОВКА ПРОБЛЕМЫ

Жизненный цикл нашей модели начинается с определения проблемы для бизнеса. Мы будем рассматривать все этапы цикла на одном примере.

Микрокредитная организация решила автоматизировать процесс выдачи кредитов и обратилась за помощью к специалистам по машинному обучению. Вместе с аналитиками

компания выяснила, что им необходима прогностическая модель, которая определяла бы по данным клиента его дефолтность.

Дефолт — невыполнение обязательств договора займа, например неоплата процентов или основного долга в установленный период.

После того как задача была поставлена, начинается процесс построения модели.

## 1. СБОР ДАННЫХ

Модель машинного обучения напрямую зависит от количества и качества передаваемых в неё данных. Чтобы получить в итоговой модели более качественные данные, на этапе сбора нам необходимо собрать максимальное их количество.

Дата-аналитики микрокредитной организации запросили у сотового оператора информацию о перемещениях клиента. Так компания сможет проверить, верно ли клиент указал место проживания и работы или предоставил ложную информацию, что, по мнению аналитиков, может стать полезным признаком.

## 2. ОЧИСТКА ДАННЫХ

После сбора данных мы должны избавиться от «мусора» (некачественных данных), и именно от этого этапа будет зависеть качество финальной модели машинного обучения.

*Garbage in — garbage out (GIGO)*, в пер. с англ. «мусор на входе — мусор на выходе».

На этапе очистки данных мы определяем пропущенные значения, аномалии и выбросы в данных.

Данные с пропусками в большинстве случаев нельзя передать в модель.

В случае с микрокредитной организацией на ранних этапах работы компании в анкете клиента не заполнялся регион проживания. Аналитики пытаются по адресу заполнить значения, а там, где это невозможно, просто проставляют значение *'unknown'*, как делали это вы в модуле *PYTHON-14*.  
*Очистка данных.*

Дублирующиеся данные и выбросы, аномальные значения не приносят пользу алгоритму.

Слишком высокие значения зарплат чаще всего являются ложной информацией среди заемщиков и могут сбивать алгоритмы с толку, поэтому для таких признаков проводят обработку выбросов. Эти действия вы выполняли в модуле *PYTHON-14*. *Очистка данных*, тем самым подготавливая почву для последующего анализа данных.

### 3. РАЗВЕДЫВАТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

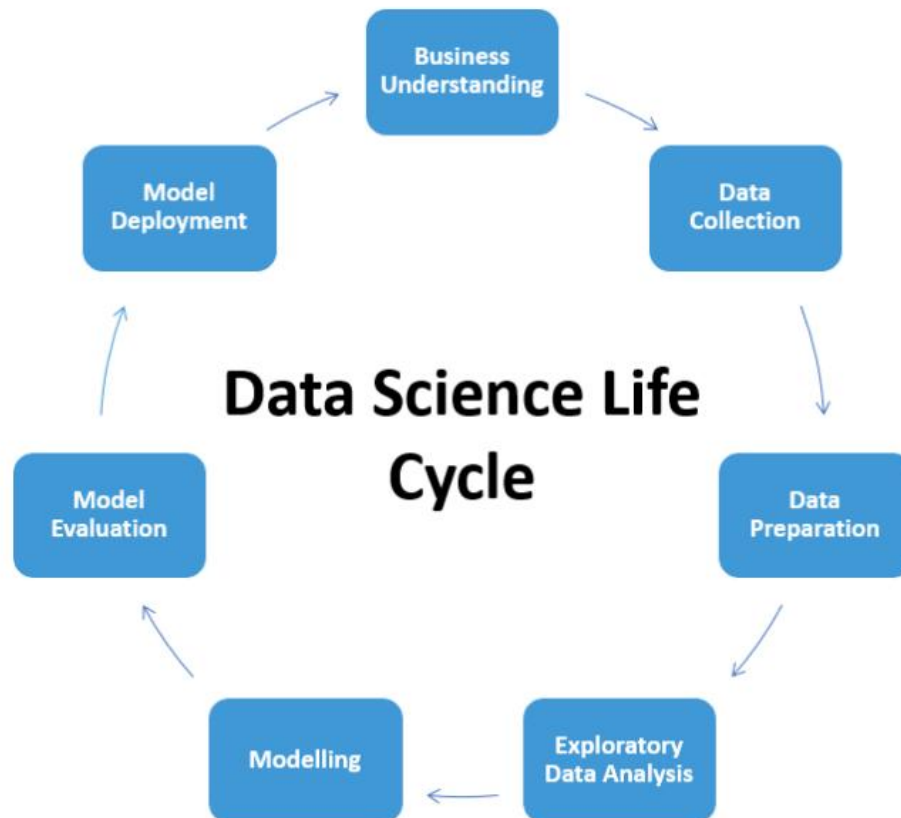
Данные ценны не сами по себе, а только если приносят пользу. Разведывательный анализ данных похож на работу детектива: имея под рукой данные, вы ищете подсказки и идеи для извлечения из них полезной информации для бизнес-задачи, которую вы пытаетесь решить. Вы исследуете одну переменную, затем две переменные вместе, а затем сразу несколько переменных. Если ваша модель показывает неудовлетворительные результаты, то почти всегда виновником является набор обучающих данных, содержащий ошибки. Они и создают потолок качества для вашей модели.

Дата-сайентисты обнаружили, что очень важным параметром является балл из кредитной истории заёмщика. Он действительно содержит в себе очень много информации о поведении заёмщика в предыдущих кредитах. Такая информация будет скорее всего полезна модели. Специалисты по данным приняли решение раскрыть этот важный признак на несколько составляющих, поскольку полагают, что новые, более сильные признаки сделают модель лучше.

## 4. ЭТАП МОДЕЛИРОВАНИЯ И ВЫВОДА МОДЕЛИ В ПРОДАКШЕН

На этом этапе ведётся построение предсказательных моделей, проверка качества и публикация модели. Эти этапы вы будете подробно изучать далее в курсе в блоках по машинному обучению.

**! Важно!** На самом деле цикл машинного обучения редко имеет линейную структуру. Он представляет собой повторяющийся цикл улучшения данных и модели. В любой момент мы можем вернуться на этап разведывательного анализа данных или даже на этап формулирования бизнес-проблемы, если качество модели неудовлетворительное.



*Цикл улучшения модели и данных. (Определение проблемы бизнеса, сбор данных, подготовка данных, разведывательный анализ данных, разработка и оценка качества модели, внедрение).*

А теперь про работу...

В реальной жизни не всегда полный цикл проводится одним человеком. На данный момент мы можем выделить три направления в машинном обучении.

① **Дата-аналитики** проводят предварительный анализ данных и их очистку, ищут закономерности в данных и ответы на запросы бизнеса.

Предварительный анализ и очистку данных вы выполняли в модуле *PYTHON-14. Очистка данных и в модулях и PYTHON-11-12. Базовые и продвинутые приёмы работы с данными в Pandas*. Определению

проблем бизнеса вы научитесь позже в курсе в блоках машинного обучения в бизнесе.

② **Дата-сайентист** работает с большими объёмами данных, создаёт рекомендательные алгоритмы и прогнозные модели. Эти этапы вы будете подробно изучать далее в курсе в блоках по машинному обучению.

③ **Дата-инженер** обеспечивает сбор и хранение данных. Дата-инженер также занимается внедрением финальной модели машинного обучения в работающую систему.

В нашем примере дата-аналитики подумали, что хорошо было бы собирать информацию об устройстве пользователя, поскольку это может оказаться важным признаком для формирования списка личных рекомендаций. Дата-инженеры ставят задачу и с этого момента в базу данных будет записываться информация об устройстве каждого пользователя. Через некоторое время её можно будет анализировать и использовать в моделях.

→ Эти этапы вы будете подробно изучать далее в курсе в блоках по инжинирингу данных.

Однако зачастую границы очень размыты: аналитик, например, также может заниматься сбором данных, а дата-сайентист — строить визуализацию и искать закономерности в данных.

Это часто зависит от размера компании. В небольших компаниях вы скорее всего будете совмещать сразу несколько направлений. В больших корпорациях за эти направления могут отвечать целые отделы.



### 3. Алгоритм и методы EDA

→ В прошлом юните мы узнали, какую роль занимает разведывательный анализ данных в жизненном цикле модели машинного обучения. В этом юните мы подробно рассмотрим, что включает в себя разведывательный анализ данных.

Недостаточно просто посмотреть на столбец или на всю таблицу и определить важные моменты в данных. Для этого были разработаны методы *EDA*, помогающие раскрыть смысл в данных. Эти методы могут выполняться в любом порядке или могут быть опущены по усмотрению специалиста.


Например, наша модель машинного обучения предсказывает рейтинг ресторанов. Мы хотим улучшить её результат. Из введения в разведывательный анализ данных мы знаем, что качественный *EDA* сможет улучшить и качество модели.

Давайте посмотрим, какие методы *EDA* нам необходимо применить.

#### **МЕТОД FEATURE ENGINEERING (ПРОЕКТИРОВАНИЕ ПРИЗНАКОВ)**

Это процесс создания новых признаков для повышения качества прогнозной модели. Для генерации новых признаков мы можем использовать внешние источники данных или конструировать признаки из имеющихся признаков в наборе данных.

**Цель этого метода** — создать новые, более сильные признаки для обучения модели. Изучать проектирование признаков мы будем далее — в модулях, посвящённых проектированию признаков.



У нас есть информация, в каком городе находится ресторан. Мы можем дополнительно создать признак из существующего, который бы говорил нам о том, является ли этот город столицей. Скорее всего, такой признак будет очень важен в определении рейтинга ресторана, ведь в столице рейтинг ресторанов в среднем выше. Сделать такой вывод нам помогает **статистический анализ**, который мы будем изучать далее в модулях про статистику.

Или, например, мы можем в открытых источниках данных найти информацию о населении этого города и создать новый числовой признак «население города, в котором находится ресторан». На первый взгляд он может оказаться действительно важным, но проверить это мы сможем только на этапе отбора признаков.

## МЕТОД FEATURE SELECTION (ОТБОР ПРИЗНАКОВ)

Это процесс выбора признаков из общего набора данных признаков, больше всего влияющих на качество модели.

Например, признак «город ресторана», из которого мы создавали новые признаки, кажется излишним и не таким важным.

Проверить это помогают различные **статистические тесты значимости**, которые мы подробно будем изучать далее в модулях про разведывательный анализ данных.

## МЕТОД КОДИРОВАНИЯ ПРИЗНАКОВ

Чаще всего в кодировании нуждаются категориальные признаки. Вы с ними познакомились в *PYTHON-11. Базовые приемы работы с данными в Pandas Юнит 5. Тип данных Category*. Они представлены

обычно в строковом формате, а большинство алгоритмов машинного обучения требуют численного формата.

У нас есть признак, говорящий о том, к какой кухне мира относят ресторан, например к средиземноморской.

В таком виде мы не можем передать данные в модель, поэтому закодируем их таким образом:

1 — средиземноморская кухня

2 — китайская

3 — грузинская

Есть и другое кодирование: когда мы переходим от численного признака к категории за счёт кодирования интервалов одним значением. Например, средний чек в ресторане мы можем закодировать как:

1 — до 1 000 руб.

2 — от 1 000 руб. до 3 000 руб.

3 — от 3 000 руб. и выше.

Такое кодирование помогает отнести ресторан к определённому ценовому сегменту. Мы можем предположить, что рестораны из высокого ценового сегмента (значение 3) скорее получат высокий рейтинг. Однако, чтобы проверить эту гипотезу, нам необходимы знания математической статистики.

Проверка статистических гипотез — также один из важных методов разведывательного анализа данных.

## 4. Знакомство с данными: винные обзоры

→ Дальнейший разбор инструментов и практические задания мы будем проводить на наборе данных из соревнования на *kaggle*.

 [Скачать датасет](#)

После просмотра документального фильма о сомелье вы захотели создать прогностическую модель для оценки вин вслепую, как это делает сомелье.

Определив бизнес-задачу, вы перешли к сбору данных для обучения модели. После нескольких недель парсинга сайта WineEnthusiast вам удалось собрать около 130 тысяч строк обзоров вин для анализа и обучения.

Вот какие признаки вам удалось собрать:

*country* — страна-производитель вина.

*description* — подробное описание.

*designation* — название виноградника, где выращивают виноград для вина.

*points* — баллы, которыми *WineEnthusiast* оценил вино по шкале от 1 до 100.

*price* — стоимость бутылки вина.

*province* — провинция или штат.

*region\_1* — винодельческий район в провинции или штате (например Напа).

*region\_2* — конкретный регион. Иногда в пределах винодельческой зоны указываются более конкретные регионы (например Резерфорд в долине Напа), но это значение может быть пустым.

*taster\_name* — имя сомелье.

*taster\_twitter\_handle* — твиттер сомелье.

*title* — название вина, которое часто содержит год и другую подробную информацию.

*variety* — сорт винограда, из которого изготовлено вино (например Пино Нуар).

*winery* — винодельня, которая производила вино.

Прочитаем наш файл с винными обзорами:

```
data = pd.read_csv('wine.csv')
```

Итак, файл прочитан. Далее мы займёмся анализом структуры собранных данных. А сейчас предлагаем вам ответить на несколько вопросов по содержимому набора данных.

## ★ 6. БОНУС. EDA одной строкой кода

→ В реальной жизни большую часть работы над моделью занимает разведывательный анализ данных. С целью решения подобной проблемы появились инструменты автоматической визуализации и представления датасета.

К таким инструментам можно отнести следующие библиотеки *Python*, которые могут выполнять *EDA* всего одной строкой кода:

- *d-tale*;
- *pandas-profiling*;
- *sweetviz*.

### PANDAS-PROFILING

→ [Ссылка на библиотеку](#)

*Pandas-profiling* — это библиотека с открытым исходным кодом, которая создаёт подробный отчёт по данным. *Pandas-profiling* можно легко использовать для больших наборов данных: отчёты создаются всего за несколько секунд.

Установка:

```
pip install pandas-profiling
```

### РАБОТА С PANDAS-PROFILING НА ПРИМЕРЕ ВИННЫХ ОБЗОРОВ

Начните свою работу с загрузки датасета винных обзоров, с которым мы познакомились в *Юните 5. Проверка*.

```
import pandas as pd
from pandas_profiling import ProfileReport

df = pd.read_csv('wine.csv')
```

Сгенерируйте отчёт одной строкой.

```
profile = ProfileReport(df, title="Wine Pandas Profiling
Report")
profile
```

→ У вас должен получиться такой отчёт.

## СКРИНКАСТ: РАБОТА С БИБЛИОТЕКОЙ PANDAS-PROFILING НА ПРИМЕРЕ ВИННЫХ ОБЗОРОВ



Результат отчёта *pandas-profiling* можно использовать для первичной визуализации набора данных, числового и статистического анализа переменных, с которым вы познакомитесь в следующих моделях по статистике. Также отчёт используется для вывода корреляции переменных, выявления пропущенных значений, а также для просмотра первых и последних строк набора данных.

# SWEETVIZ

→ [Ссылка на библиотеку](#)

*Sweetviz* — это библиотека автоматического анализа с открытым исходным кодом. *Sweetviz* также можно использовать для сравнения нескольких наборов данных и выводов по ним. Это может быть удобно, когда необходимо сравнить обучающий и тестовый наборы данных. Об этом вы узнаете далее в модулях про машинное обучение.

Установка:

```
pip install sweetviz
```

С помощью *Sweetviz* можно провести первичный осмотр набора данных, просмотреть свойства признаков, провести числовой и статистический однофакторный анализ и так далее.

## РАБОТА СО SWEETVIZ НА ПРИМЕРЕ ВИННЫХ ОБЗОРОВ

Начните свою работу с загрузки датасета винных обзоров.

```
import pandas as pd
import sweetviz as sv

df = pd.read_csv('wine.csv')
```

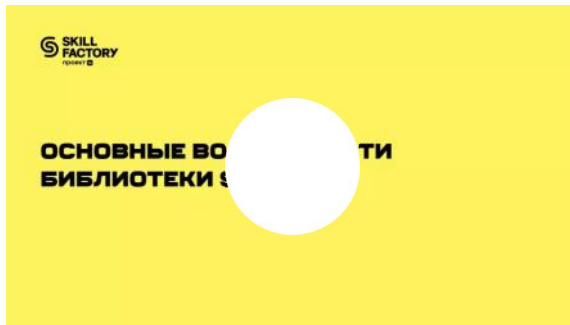
Чтобы проанализировать набор данных, просто используйте функцию `analyze()`, а затем — `show_html()`:

```
report = sv.analyze(my_dataframe)
report.show_html()
```

У вас должен получиться [такой отчёт](#).



## СКРИНКАСТ: РАБОТА С БИБЛИОТЕКОЙ SWEETVIZ НА ПРИМЕРЕ ВИННЫХ ОБЗОРОВ



### D-TALE

→ [Ссылка на библиотеку](#)

*D-Tale* — это библиотека с открытым исходным кодом. *D-Tale* делает подробный разведывательный анализ набора данных. Интересная особенность: библиотека предоставляет функцию экспорта кода для каждого графика или элемента анализа в отчёте.

Установка:

```
pip install dtale
```

### РАБОТА С D-TALE НА ПРИМЕРЕ ВИННЫХ ОБЗОРОВ

Начните свою работу с загрузки датасета винных обзоров.

```
import pandas as pd
import dtale

df = pd.read_csv('wine.csv')
```

Сгенерируйте отчёт одной строкой.

```
d = dtale.show(df)
d
```

У вас должен получиться такой отчёт.

## СКРИНКАСТ: РАБОТА С БИБЛИОТЕКОЙ D-TALE НА ПРИМЕРЕ ВИННЫХ ОБЗОРОВ

### СРАВНЕНИЕ БИБЛИОТЕК

Основные возможности/библиотека	Pandas- profiling	Sweetviz	D-Tale
---------------------------------	----------------------	----------	--------

Тип, уникальные значения, пропущенные значения, повторяющиеся строки, наиболее частые значения	+	+	+
Числовой анализ (мин, макс, квартили, мода/медиана...)	+	+	+
Описательная статистика	+	+	+
Визуализация	+	+	+
Сравнение различных наборов данных	-	+	+
Экспорт кода	-	-	+

**Совет.** Мы советуем проводить самостоятельный разведывательный анализ данных, используя библиотеку *pandas* и писать код на *Python*, так как важно знать и применять основы *EDA*. Помните, что ни одна из перечисленных выше библиотек не сможет провести детальный анализ так, как это делает специалист по данным. Данные библиотеки можно использовать для ускорения вашей работы, первичной визуализации данных, например чтобы понять, с какими данными вам придётся работать.

↓ Прежде чем перейти к следующему юниту, ответьте на вопросы для закрепления знаний.

→ В этом юните мы изучили инструменты для ускорения процесса *EDA*, а в следующем подведём итоги.