

Финальная работа

Анализ сайта «СберАвтоподписка»

Содержание

О компании «СберАвтоподписка»	3
О продукте	6
Что нужно сделать	10
Глоссарий	11
План работы	12
Формат решения	15
Критерии оценки	16
Описание данных	19
Полезные материалы для выполнения проекта	20

Поздравляем вас, вы дошли до финальной темы нашего курса!

В рамках этой финальной работы вы продемонстрируете свои знания и навыки для решения реальной задачи от нашего партнёра — «СберАвтоподписки».

Реальные задачи не похожи на учебные из-за того, что в них нет простых правильных ответов. Вы станете профессионалом только тогда, когда сможете научиться использовать учебные и модельные подходы к решению реальных задач с ошибками в данных, неоднозначных проблем и находить непростые решения.

О компании «СберАвтоподписка»

«СберАвтоподписка» — это сервис долгосрочной аренды автомобилей для физлиц.

Клиент платит фиксированный ежемесячный платёж и получает в пользование машину на срок от шести месяцев до трёх лет. Также в платёж включены:

- страхование (КАСКО, ОСАГО, ДСАГО);
- техническое обслуживание и ремонт;
- сезонная смена шин и их хранение;
- круглосуточная служба поддержки.

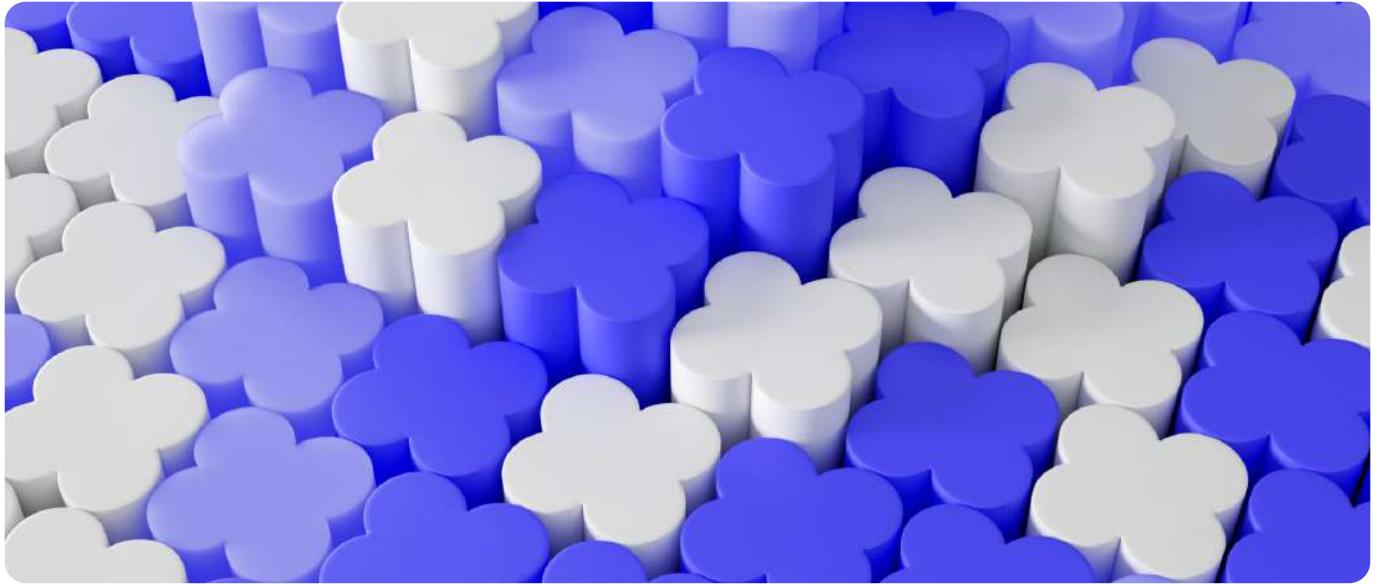
За дополнительную сумму можно приобрести услугу консьерж-сервиса — доставку автомобиля до сервисного центра и обратно на техническое обслуживание, сезонную замену шин, ремонт.

Один год исполнился «СберАвтоподписке» 24 мая 2022 года.

«СберАвтоподписка» предлагает новый для российского рынка способ владения автомобилем и выступает в качестве альтернативы автокредиту. Машина — это актив, который быстро теряет в цене, поэтому мы рекомендуем вкладывать собственные средства в финансовые инструменты, а автомобиль брать во временное пользование за комфортную сумму в месяц. Кроме этого, «Автоподписка» обладает преимуществами перед покупкой машины в кредит и использованием каршеринга:

	«СберАвтоподписка»	Покупка машины	Кредит	Каршеринг
Нет крупного первого платежа	+	—	+	
Ежемесячный платёж		Сравним с кредитом на срок от пяти лет	+	Оплата за время поездки
Не надо платить за страхование и обслуживание	+	—	+	
Постоянная доступность	+	+		—
Индивидуальное использование	+	+		—
Гарантированная исправность машины	+	—		—
Отсутствие промонаклеек	+	+		—

Подробнее об отличиях подписки от каршеринга по [здесь](#).



“

За год работы нам удалось доказать востребованность сервиса. Более 1600 довольных клиентов и более 1700 автомобилей в подписку — некоторые клиенты взяли в пользование более одного автомобиля. Эти цифры — лучшее подтверждение того, что подписка — не просто дань моде.

Несмотря на то, что последние годы были трудными для автобизнеса, нам удалось предоставить клиентам выгодные условия. Сейчас мы видим рост интереса к услуге подписки. Многие не готовы к покупке автомобиля по новым ценам, ожидая их снижения в будущем, и выбирают наш продукт как доступную альтернативу владения.

Никита Чернов, коммерческий директор сервиса
«СберАвтоподписка»

О продукте

В каталоге сервиса представлены более 20 моделей в наличии. Также сервис практикует покупку машины под конкретного пользователя для тех, кто готов взять автомобиль в подписку, но не нашёл на сайте интересующий вариант.

Популярные автомобили

LADA Granta



В месяц от
27 700 ₽

Renault Logan Stepway



В месяц от
38 250 ₽

Skoda Karoq



В месяц от
49 100 ₽

Хочу заказать другой автомобиль

Оставить запрос

Haval Jolion



В месяц от

Chery Tiggo 4



В месяц от

Geely Coolray



В месяц от

Чтобы оформить подписку, необходимо кликнуть по понравившемуся автомобилю и выбрать:

- срок аренды: от шести месяцев до трёх лет;
- пакет пробега: лимит по количеству пройденных километров.

Skoda Karoq

2021

на срок:

Ежемесячно | 24 мес. | **36 мес.**

Пробег на весь срок подписан:

30000 км | 30000 км | 30000 км

Приемка

49 100 ₽

Каждый месяц

Оформить подписку

Входит в подписку

После клика по кнопке «Оформить подписку» нужно оставить свои контактные данные.

СБЕР АВТОПОДПИСКА

Оставьте контактные данные

Ваш номер телефона

E-mail

Пароль

Согласие на обработку персональных данных на [Сбербанк Онлайн](#)

Жду звонка

Если звонок не будет, звоните **8 800 663 63 63**

Приём за отказ при подписке:

3000 км | 30000 км | 30000 км

Страховка:

ОСАГО + КАСКО (без франшизы)

Есть практик?

Подождите практик

35 100 ₽/месяц

Поддержка 24/7

Техобслуживание

Страховки

Сервисное сопровождение

Налоги

Включено

Включено

Включено

Включено

Включено

Оставить заявку

На сайте пользователь совершает целевые действия. Например, нажимает кнопки типа «Оставить заявку», «Заказать звонок».

Или нецелевые действия, например, просмотр карточек авто или «блуждания» по основной странице и страницам с помощью.

Когда заявка оставлена, с пользователем связывается менеджер для подбора оптимального варианта подписки. Клиент предоставляет документы и «СберАвтоподписка» проводит скоринг, в результате чего принимает решение, готовы ли мы выдать клиенту автомобиль в подписку. После успешного скоринга подбор автомобиля продолжается, в том числе возможен вариант покупки конкретного автомобиля специально под клиента. После оплаты счёта клиент получает автомобиль в подписку: его можно забрать в Москве, а также возможна доставка автомобиля в другие города России.

После оформления подписки пользователь взаимодействует со «СберАвтоподпиской» по телефону и с помощью электронной почты. Ссылки на оплату услуги приходят по почте и SMS.

История проекта



Александр Тимаков

Руководитель отдела анализа данных

Привет! Мы — команда аналитики „СберАвтоподписки“. Через наши руки проходит множество разных данных, большая их часть — из различных внутренних систем и с нашего веб-сайта. Мы стараемся собирать многие события с сайта, чтобы извлекать полезные инсайты и на их основе проводить работу, которая помогает нашему бизнесу.

Недавно к нам пришла продуктовая команда и начала задавать вопросы про веб-сайт и поведение пользователей на нём. Они хотят понять, из каких источников к нам идёт самый полезный трафик, какие авто пользуются наибольшим спросом, с каких устройств чаще всего заходят на сайт, какие действия и формулировки вызывают наибольший отклик — и массу других вещей.

Вам нужно будет разобраться в данных, поступающих с веб-сайта, и помочь продуктовой команде ответить на ряд вопросов — как на общие (о характеристиках пользователей и общей статистике по посещаемости), так и на более комплексные (о тестировании гипотез, разработке предиктивных моделей и разработке пайплайнов обработки данных).

Что нужно сделать

Итак, **ваша задача** — изучить предоставленный датасет, ответить на вопросы из общей части (подразумевающей базовую обработку данных и их разведочный анализ) и выполнить задание по специализации, в которой вы хотели бы развиваться в будущем.

Шаги по каждому пункту и вопросы будут дальше, в разделе «План работы».

В финальной работе представлены задачи для трёх специализаций.

1 Дата аналитик:

Разбирается в математической статистике, умеет формулировать и исследовать гипотезы. Владеет инструментами Business Intelligence, которые помогают исследовать данные и переводить информацию, которая в них содержится, на язык бизнеса, понятный людям, которые принимают то или иное управленческое решение.

2 ML-инженер:

Разбирается в математической статистике, умеет формулировать и исследовать гипотезы. Владеет инструментами Business Intelligence, которые помогают исследовать данные и переводить информацию, которая в них содержится, на язык бизнеса, понятный людям, которые принимают то или иное управленческое решение.

3 Дата-инженер:

программист, который разбирается в специфике сбора, хранения и обработки больших и не очень данных. Он знает, как правильно собрать и куда положить данные для различных задач. Кроме того, он внедряет сервисы, которые разрабатывают ML-инженеры, ставит их на мониторинг, а именно: следит за тем, чтобы они правильно работали в любой момент времени.

Глоссарий

Целевое действие — события типа «Оставить заявку» и «Заказать звонок»
(ga_hits.event_action in ['sub_car_claim_click', 'sub_car_claim_submit_click',
'sub_open_dialog_click', 'sub_custom_question_submit_click',
'sub_call_number_click', 'sub_callback_submit_click', 'sub_submit_success',
'sub_car_request_submit_click']).

CR (Conversion Rate) — показатель конверсии из визита (уникальный session_id) в любое целевое действие в рамках одного визита (в случае наличия >1 целевого действия — считать все как одно).

Органический трафик — все визиты с ga_sessions.utm_medium in ('organic', 'referral', '(none)').

Платный трафик — весь неорганический трафик.

Информация про марку и модель авто — содержится в ga_hits.hit_page_path.

Реклама в социальных сетях — все визиты с ga_sessions.utm_source in ('QxAxdyPLuQMEcrdZWdWb', 'MvfHsxITijuriZxsqZqt', 'ISrKoXQCxqqYvAZICvjs',
'IZEXUFLARCUMynmHNBGo', 'PlbkrSYoHuZBWfYjYnfw',
'gVRrcxiDQuJiljoTbGm').



План работы

Проведите подготовительную работу (1 час):

- Прочитайте предоставленный датасет.
- Ознакомьтесь с описаниями представленных атрибутов.
- Оцените полноту и чистоту данных. Попытайтесь понять, что стоит за этими данными в реальном мире. Приведите данные в удобный / нормальный вид для дальнейшей работы.

Проведите разведочный анализ данных (4 часа):

- Проведите базовую чистку (дубликаты, пустые значения, типизация данных, ненужные атрибуты).
- Посмотрите на распределение ключевых атрибутов, их отношения.

Выполните задание согласно вашей специализации (18 часов):

DA:

Проведите проверку следующих гипотез:

- Органический трафик не отличается от платного с точки зрения CR (Conversion Rate) в целевые события.
- Трафик с мобильных устройств не отличается от трафика с десктопных устройств с точки зрения CR (Conversion Rate) в целевые события.
- Трафик из городов присутствия (Москва и область, Санкт-Петербург) не отличается от трафика из иных регионов с точки зрения CR (Conversion Rate) в целевые события.

Дайте ответы на вопросы продуктовой команды:

- Из каких источников / кампаний / устройств / локаций к нам идёт самый целевой трафик (и с точки зрения объёма трафика, и с точки зрения CR)?
- Какие авто пользуются наибольшим спросом? У каких авто самый лучший показатель CR (Conversion Rate) в целевые события?
- Стоит ли нам увеличивать своё присутствие в соцсетях и давать там больше рекламы?

ML:

- Научитесь предсказывать совершение целевого действия (ориентировочное значение ROC-AUC ~ 0.65) — факт совершения пользователем целевого действия.
- Упакуйте получившуюся модель в сервис, который будет брать на вход все атрибуты, типа `utm_*`, `device_*`, `geo_*`, и отдавать на выход 0/1 (1 — если пользователь совершил любое целевое действие).

DE:

- Настройте и запустите локальную БД, подходящую для хранения и исполнения запросов к данным в предоставленном датасете.
- Создайте объекты в БД для хранения данных исходного файла.
- Обработайте и поместите в БД данные из предоставленного основного датасета.
- Настройте пайплайн сбора, обработки и записи в БД новых .json-файлов.

Подготовьте получившиеся артефакты, проверьте их на соответствие итоговым критериям оценки и сдайте их (2 часа).

Формат решения

1 Общее:

- Формат файла — Jupyter Notebook
- Содержание — код для чтения, обработки и EDA; получившиеся графики и текстовые комментарии, а также выводы по данным.

2 DA:

- Формат файла — Jupyter Notebook.
- Содержание — результаты проверки гипотез в формате «Гипотеза — Статистические критерий и тест — Результат теста — Текстовая интерпретация результата».

3 ML:

- Формат файла — (минимум) .py-скрипт с инструкцией по локальному запуску или (максимум) localhost web app.
- Содержание — модель, берущая на вход строку с данными по визиту (согласно схеме данных) и отдающая на выход результат предсказания по отдельному событию в числовом формате 0|1.

4 DE:

- Формат файла — (минимум) .py-скрипт с DAG'ами и DDL-скриптами или (максимум) работающий пайплайн в localhost Airflow и созданные объекты в localhost SQL DB.
- Содержание — пайплайн, на вход берущий новые json-файлы и успешно добавляющий их в существующие таблицы.

Критерии оценки

Общая часть

EDA:

- Корректная обработка данных: выброшены пустые атрибуты, убраны дубликаты, обработаны NaN-значения, таблицы сцеплены корректно.
- Наличие релевантных визуализаций — подразумевается наличие визуализаций распределений (histogram / boxplot), опционально — динамика визитов и событий во времени
- Нахождение основных корреляций (heatmap + выводы по полученным связям).

Специализации

DA:

1 Проверка гипотез. Для каждой гипотезы:

- Корректность выбранного метода.
- Корректность выбранных параметров теста (мощность и прочее).
- Корректность реализации теста на Python.

2 Ответы на вопросы бизнеса (оценка отраслевого эксперта):

- Основанность на данных — ответы являются фактически корректными и не содержат утверждений, противоречащих предоставленным данным.
- Business Acumen — ответы демонстрируют понимание основ маркетинговой аналитики и/или автомобильного рынка.
- Интерпретируемость — ответ даёт однозначное решение поставленного вопроса и не требует дополнительной трактовки или расширенного комментария

ML:

- 1 Значение ошибки на тестовой выборке (сравнение с бейслайном по идентичной метрике качества).
- 2 Подход к работе с фичами:
 - Оценка проделанной работы с проработкой фичей: есть ли инжиниринг новых фичей? Релевантны ли применяемые для инжиниринга методы? Насколько линейно зависимы итоговые фичи друг от друга?
 - Насколько интерпретируемыми являются итоговые фичи в модели: можно ли описать их суть понятным языком для человека, которому нужно будет корректировать параметры рекламных кампаний?
 - Насколько итоговые фичи значимы для результата с точки зрения оказываемого фичей вклада на итоговое значение предсказываемого целевого действия?
- 3 Выбор модели (релевантность для решения поставленной задачи).
- 4 Реализация API:
 - Масштабируемость.
 - Корректность и скорость работы (среднее время от отправки запроса в API до получения ответа клиентом не превышает трёх секунд).

DE:

- 1 Корректность создания объектов в БД:
 - Корректная типизация данных.
 - Настройка первичных и внешних ключей.

2 Корректность добавления новых данных в БД:

- Обработка и типизация новых данных.
- Имеется механизм разрешения конфликтов при записи.

3 Реализация Airflow-пайплайна:

- Наличие и корректность всех зависимостей между DAG'ами / task'ами
- Качество реализации (чистота кода, идемпотентность, использование переменных, наличие обращений в БД вне task'ов).



Описание данных

Данные из Google Analytics (last-click attribution model) по сайту «СберАвтоподписка».

1 GA Sessions (ga_sessions.pkl)

Одна строка = один визит на сайт.

Описание атрибутов:

- session_id — ID визита;
- client_id — ID посетителя;
- visit_date — дата визита;
- visit_time — время визита;
- visit_number — порядковый номер визита клиента;
- utm_source — канал привлечения;
- utm_medium — тип привлечения;
- utm_campaign — рекламная кампания;
- utm_keyword — ключевое слово;
- device_category — тип устройства;
- device_os — ОС устройства;
- device_brand — марка устройства;
- device_model — модель устройства;
- device_screen_resolution — разрешение экрана;

- device_brand — марка устройства;
- device_model — модель устройства;
- device_screen_resolution — разрешение экрана;
- device_browser — браузер;
- geo_country — страна;
- geo_city — город.

1 GA Hits (ga_hits.pkl)

Одна строка = одно событие в рамках одного визита на сайт.

Описание атрибутов:

- session_id — ID визита;
- hit_date — дата события;
- hit_time — время события;
- hit_number — порядковый номер события в рамках сессии;
- hit_type — тип события;
- hit_referer — источник события;
- hit_page_path — страница события;
- event_category — тип действия;
- event_action — действие;
- event_label — тег действия;
- event_value — значение результата действия.

Полезные материалы для выполнения проекта. ROC-AUC

Вспомните матрицу ошибок в Python. В ней есть четыре элемента:

- **TN** (True Negative) — элементы, верно классифицированные как класс 0.
- **TP** (True Positive) — элементы, верно классифицированные как класс 1.
- **FN** (False Negative) — элементы, неверно классифицированные как класс 0.
- **FP** (False Positive) — элементы, неверно классифицированные как класс 1.

	Класс 0 / negative (pred_label)	Класс 1 / positive (pred_label)
Класс 0 negative (true_label)	TN (True Negative)	FP (False Positive)
Класс 1 positive (true_label)	FN (False Negative)	TP (True Positive)

Количество элементов класса 1,
верно предсказанных как класс 1

Введём новые понятия:

- **TPR (True positive rate)** — доля верно предсказанных классов у объектов, относящихся к положительному классу.

$$TPR = \frac{TP}{TP + FN} .$$

- **FPR (False positive Rate)** — доля неправильно предсказанных классов среди объектов отрицательного класса.

$$FPR = \frac{FP}{TN + FP}.$$

- **ROC-кривая (Receiver operating Characteristic)** — кривая, которая описывает взаимосвязь между чувствительностью модели (TPR — доля истинно положительных примеров) и её специфичностью (описываемой в отношении долей ложноположительных результатов — 1-FPR).

Предположим, у вас есть модель логистической регрессии, которая решает некоторую задачу бинарной классификации. При этом она предсказывает не только 0 и 1, но и может предсказывать вероятности отношения к классу 0 и к классу 1 с помощью метода `predict_proba`.

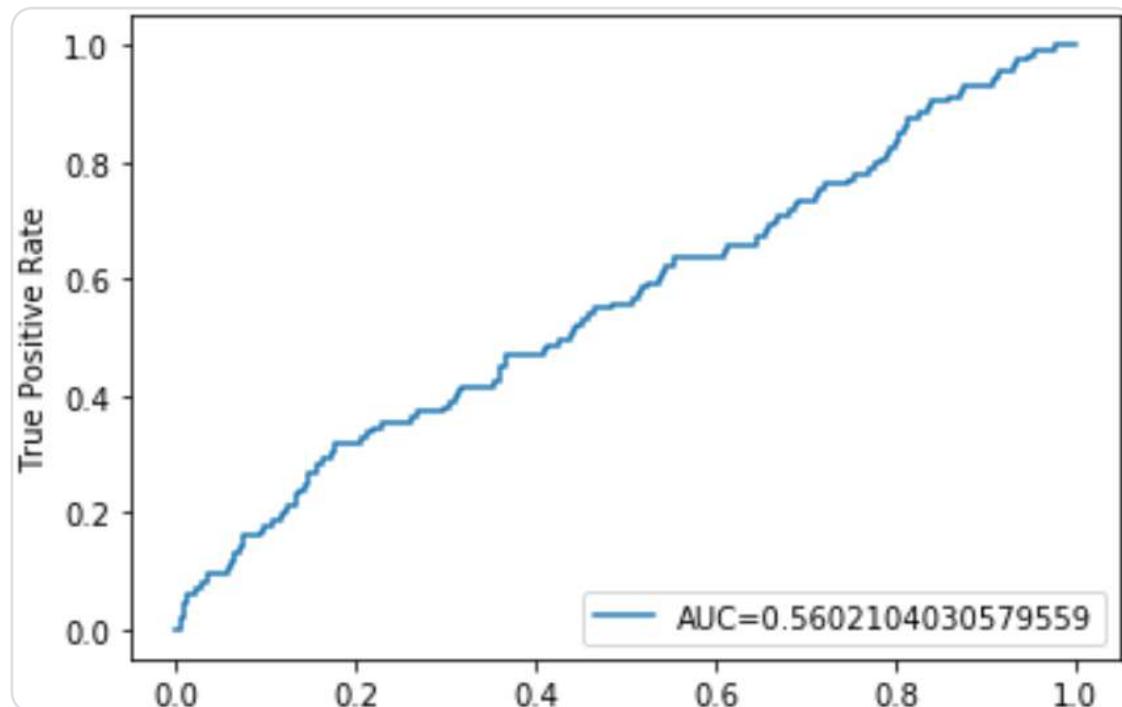
Пусть эта модель предсказала четыре вероятности отношения объекта к классу 1 (колонка `predict_proba`). Если мы возьмём стандартный порог, равный 0.5, то в результате предикта мы получим классы из колонки 2. Если же попробуем варьировать порог и выберем его равным 0.9, то получим значения из последней колонки.

true_class	<code>predict_proba</code> (class 1)	<code>predict</code> (<code>trsh=0.5</code>)	<code>predict</code> (<code>trsh=0.9</code>)
1	0.8	1	0
0	0.51	1	0
1	0.92	1	1
0	0.23	0	0

Тогда при $\text{trsh}=0.5$ TPR будет равен $2/(2+0) = 1$, а $\text{FPR}=1/(1+1)=\frac{1}{2}$.

При $\text{trsh}=0.9$ TPR будет равен $1/(1+1) = 1/2$, а $\text{FPR}=0$.

ROC-кривая — это зависимость TPR от FPR. Таким образом, мы получили две точки для ROC-кривой: $(\frac{1}{2}, 1)$ и $(0, \frac{1}{2})$. Кривая начинается в начале координат и стремится к $(1,1)$. Чем больше значений трешхолда мы исследуем, тем детальнее получается кривая.

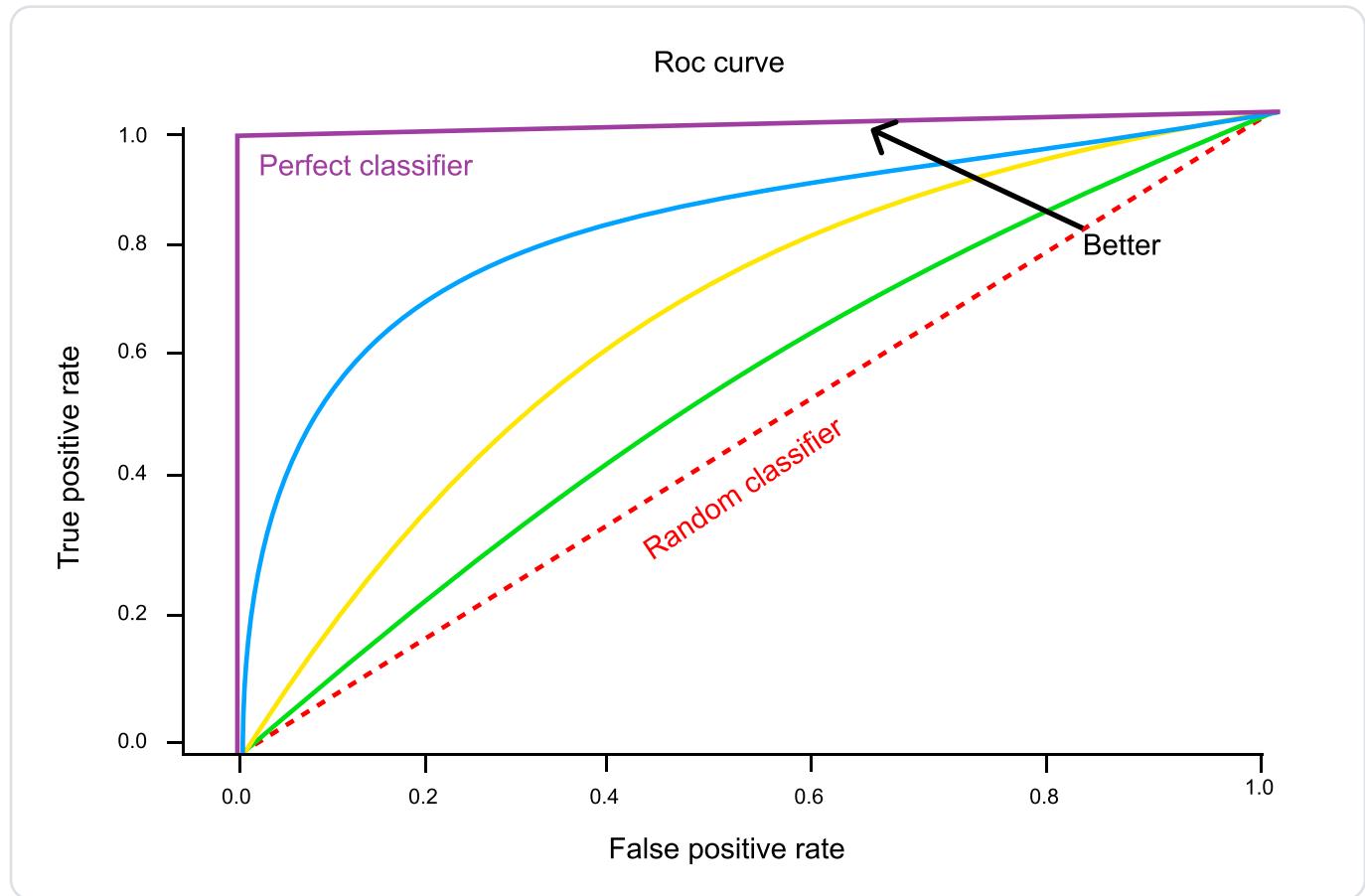


Пример ROC-кривой:

ROC-AUC (AUC = area under the curve) — одна из метрик оценки качества бинарной классификации.

Чтобы оценить, насколько хорош классификатор, рассчитывают площадь под ROC-кривой. Значение, равное единице — это идеальный классификатор. Значение 0.5 — рандомный классификатор.

То есть идеальная модель та, в которой доля TPR максимально высока, а доля FPR удерживается как можно ниже.



Значение метрики ROC-AUC вы сможете [посчитать с помощью функции](#) в sklearn, передав на вход истинные значения классов и вероятности предсказания модели.