академия больших данных

mail.ru group

MADE

Машинное обучение

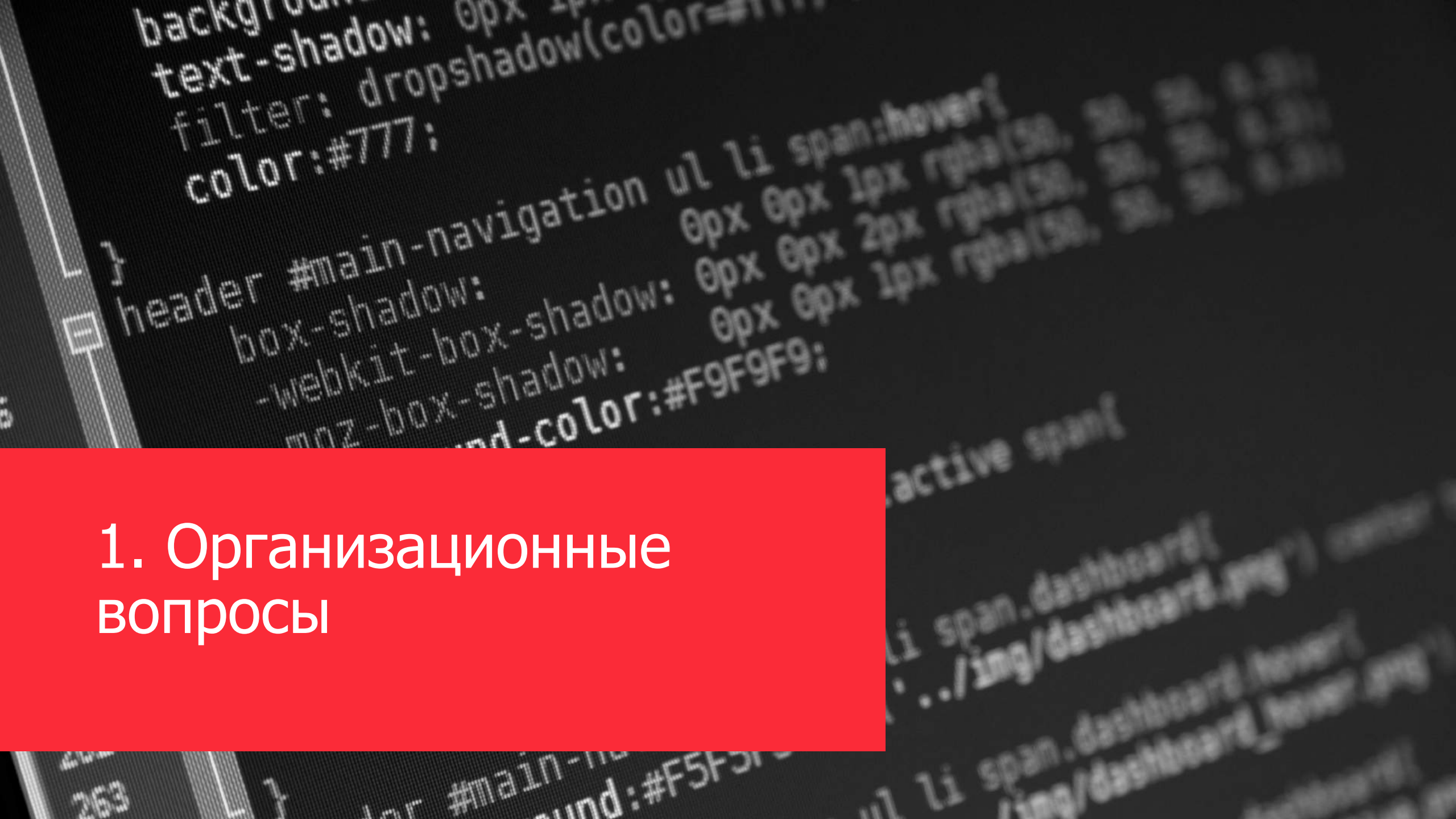
Лекция 1. Практические рекомендации и постановка задач

Кантор Виктор

Программный директор академии MADE

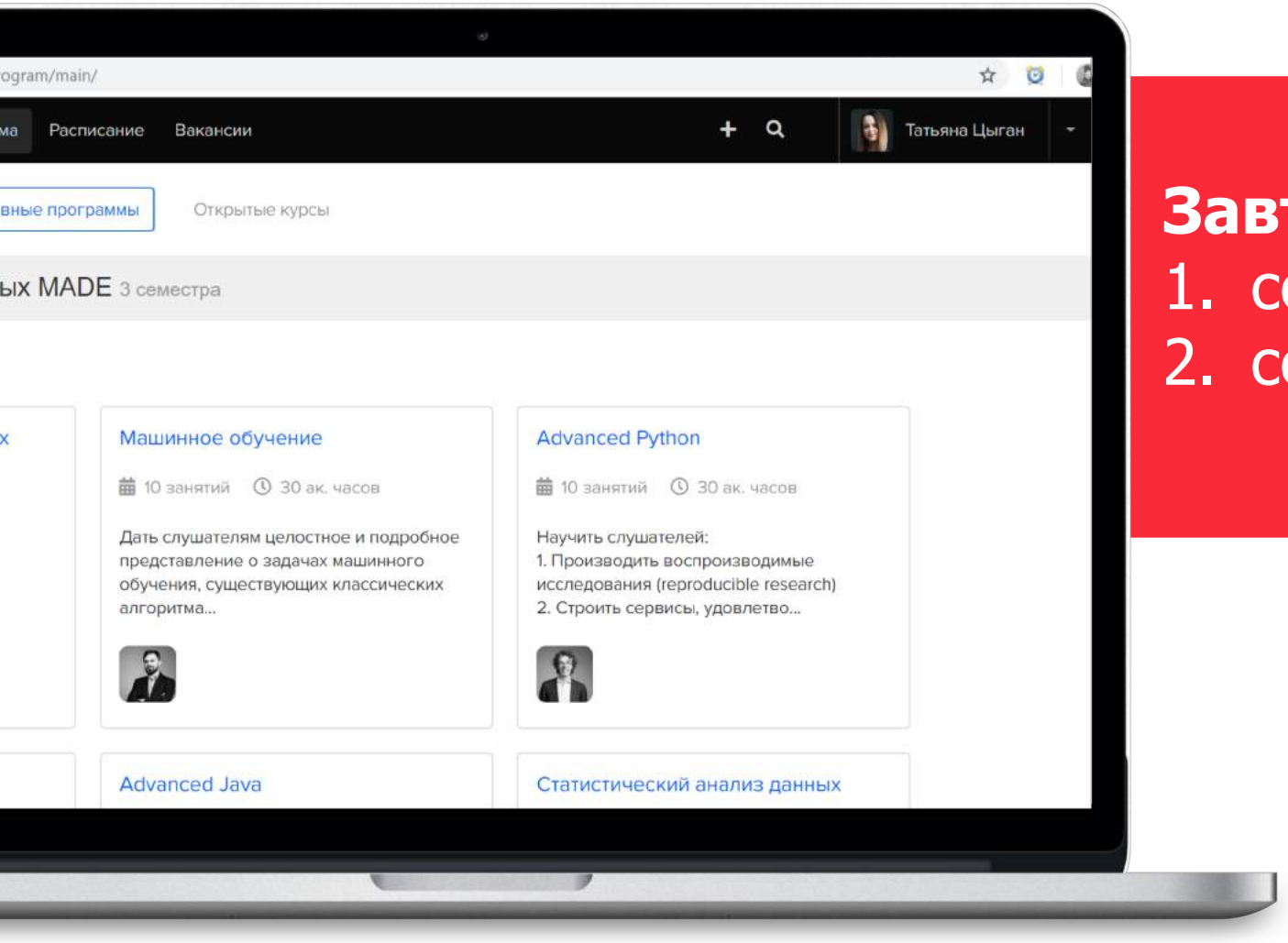
Сегодня на лекции

1. Организационные моменты
2. Задача машинного обучения
3. Постановка задач в терминах машинного обучения
4. Оценка эффекта для бизнеса



1. Организационные вопросы

Расписание, задания, материалы



Завтра вы получите:

1. ссылку на портал академии
2. ссылку на slack академии

Ближайшие занятия

9 октября

Advanced C++

12 октября

Статистический анализ данных,
Advanced Java,
Алгоритмы и структуры данных

Advanced Python начинается 15 октября

Человек, который знает всё

Татьяна Цыган
Руководитель проекта

Татьяне можно задавать любые организационные вопросы:

- в почте: t.tsygan@corp.mail.ru
- в Телеграме и слаке ODS @TatsianaTsygan
- в слаке Академии

Вопросы по программе

Виктор Кантор
Программный директор

Можно спрашивать, почему программа устроена так, какой смысл в нее вложен и какой предмет выбрать

- в Телеграме: @vkantor
- вживую в Академии

Как отчисляем из академии

- Если не присылаете первое ДЗ по машинному обучению (выдается 14 октября, дедлайн – 21 октября)
- Если к 14 ноября сдано меньше двух ДЗ суммарно по всем предметам
- Если в конце семестра не сданы два обязательных курса и один курс по выбору

Курс машинного обучения

Введение

Постановка задач машинного обучения. Библиотеки для анализа данных.

Линейные модели в классификации и регрессии

Как работают и где используются линейные модели. Задачи с разреженными признаками. Особенности использования линейных моделей: онлайн-обучение, комбинирование методов оптимизации. Hashing trick и Vowpal Wabbit. Повышение устойчивости линейных моделей.

Решающие деревья и ансамбли

Как работают решающие деревья и ансамбли деревьев из популярных библиотек (XGBoost, LightGBM, CatBoost). Построение ансамблей на практике в соревнованиях и в продакшене. Блендинг и стекинг алгоритмов.

Оценка качества

Метрики качества на исторических данных и в продакшене. Особенности оценки качества в продакшене.

Курс машинного обучения

Понижение размерности

Олдскульные методы (PCA, SVD, NMF), эмбеддинги (word2vec, glove, fasttext) и manifold learning (t-SNE).

Кластеризация

Постановка задачи и валидация кластеризации. Стандартные методы кластеризации. Кластеризация в теории и на практике на примере кластеризации текстов.

Компьютерное зрение

Задачи компьютерного зрения. От олдскульных подходов к deep learning

Обработка текстов на естественном языке и распознавание речи

Задачи NLP и Speech Recognition. От олдскульных подходов к нейросетям и deep learning

Информационный поиск

Задачи информационного поиска. Обучение ранжированию.

Работа на занятиях

На лекциях:

- Кратко обсуждаем как работают модели
- Подробно обсуждаем, где используются, как их настраивать, какие есть подводные камни и трюки на практике

На семинарах:

Разбираем интересные примеры работы моделей и решения задач

Сдача курса

Будет 4 домашних задания:

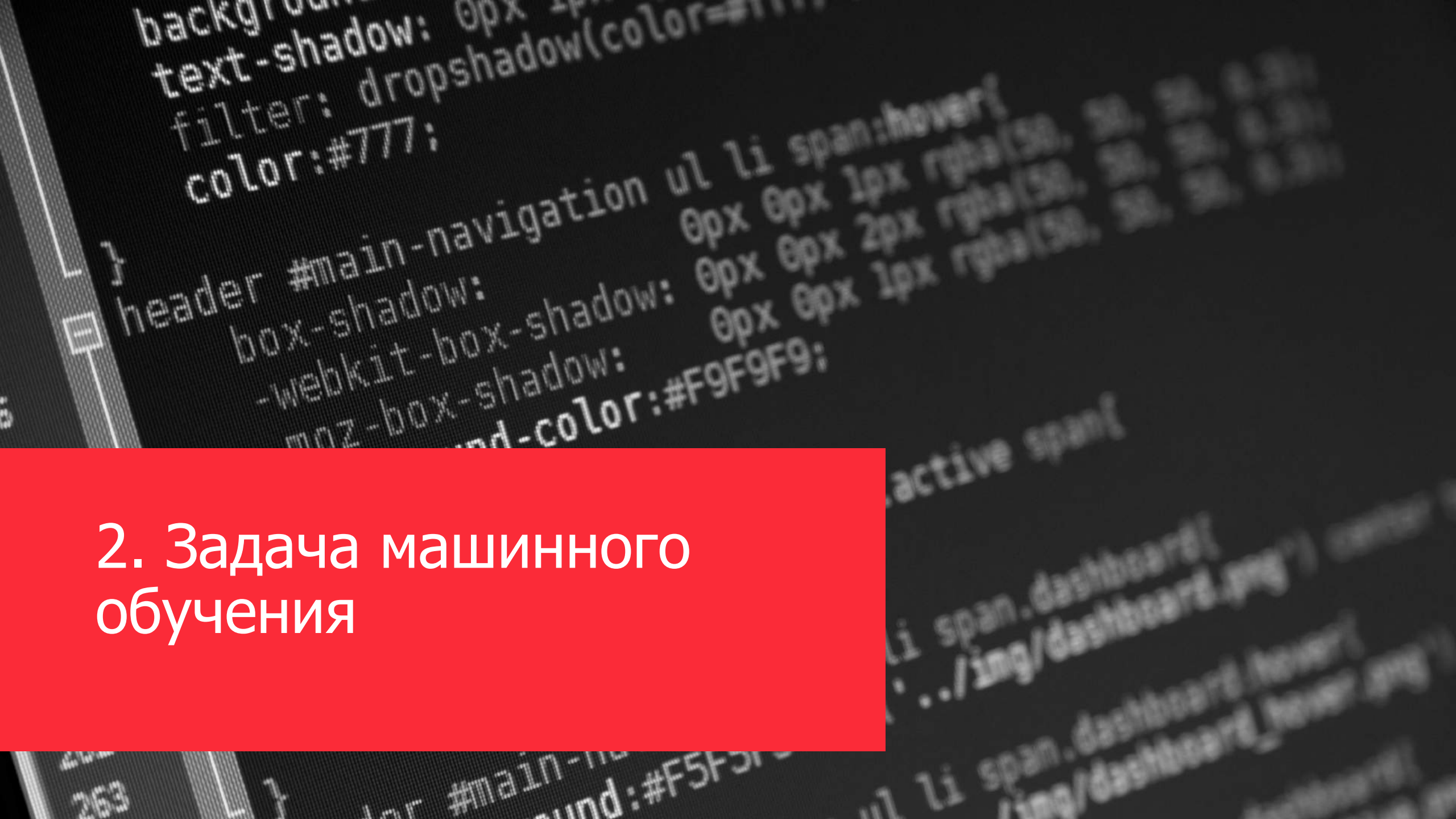
- первое со сроком выполнения одна неделя
- еще три – двухнедельные

Каждое задание оценивается по 10-балльной шкале

Итоговый балл за курс:

0.5 x ДЗ 1 + ДЗ 2 + ДЗ 3 + ДЗ 4 (максимум 35 баллов)

Также будут возможности получить дополнительные баллы



2. Задача машинного обучения

Базовая терминология

У нас есть **объекты** $x_1, x_2, \dots, x_n$ (например, пользователи),

каждый из которых описан числами, называемыми **признаками** (например, количеством кликов по баннерам разных категорий за последнюю неделю)

Т.е. каждый объект рассматривается как **вектор признаков:**

$$x_i = (x_{i1}, x_{i2}, \dots, x_{id})$$

Базовая терминология

Для этих объектов известны **целевые значения (target)** $y_1, y_2, \ldots, y_n$ (например - 1 для тех, кто кликнул по баннеру и 0 для тех, кто не кликнул)

Объекты с известными на них ответами $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ составляют **обучающую выборку**

Задача: для новых объектов $x_{n+1}, x_{n+2}, \ldots, x_{n+t}$, на которых мы не знаем целевые значения **(тестовой выборки, test set)**, спрогнозировать их

Базовая терминология

Задача: для новых объектов $x_{n+1}, x_{n+2}, \ldots, x_{n+t}$, на которых мы не знаем целевые значения **(тестовой выборки, test set)**, спрогнозировать их

Что не совсем правильно?

Базовая терминология

Задача: для новых объектов $x_{n+1}, x_{n+2}, \ldots, x_{n+t}$, на которых мы не знаем целевые значения **(тестовой выборки, test set)**, спрогнозировать их

Что не совсем правильно?

Задача – настроить параметры алгоритма, прогнозирующего целевые значения для новых объектов

Базовая терминология

Задача – настроить параметры алгоритма $a(x)$, прогнозирующего целевые значения для новых объектов

Такой алгоритм мы будем называть **моделью**

Как это выглядит

Средний заказ	Средняя оценка	Частота заказов	…	Лет в сервисе	Ушел?
984 р	2.5	4	…	0.1	**Да**
563 р	4.1	6	…	1.5	**Нет**
551 р	3.9	2	…	0.2	**Да**

Обучающая выборка

Как это выглядит

Средний заказ	Средняя оценка	Частота заказов	…	Лет в сервисе	Уйдет?
498 р	3.5	3	…	1.6	
1232 р	4.6	7	…	0.5	
757 р	2.8	6	…	3.0	

Тестовая выборка

Вопросы

1. Что значит «спрогнозировать»?

Метрики

$y_1, y_2, \dots, y_n$ - ответы (целевые значения)

$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$- прогнозы (на тех же объектах)

Примеры метрик:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}[y_i = \hat{y}_i]$$

Метрики

$y_1, y_2, \dots, y_n$ - ответы (целевые значения)

$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$- прогнозы (на тех же объектах)

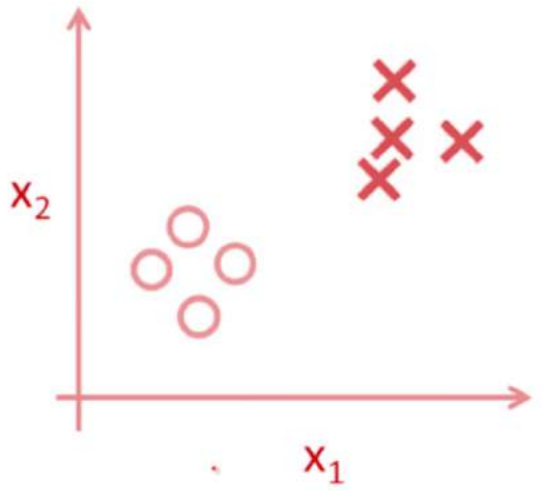
Примеры метрик:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}[y_i = \hat{y}_i]$$

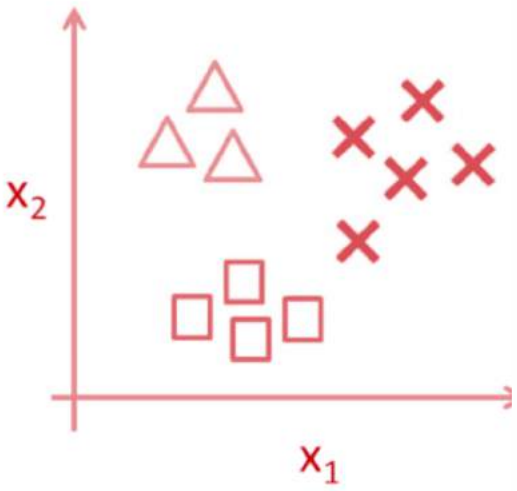
$$Precision = \frac{\sum_{i=1}^{n}[y_i = \hat{y}_i = 1]}{\sum_{i=1}^{n}[\hat{y}_i = 1]} \qquad Recall = \frac{\sum_{i=1}^{n}[y_i = \hat{y}_i = 1]}{\sum_{i=1}^{n}[y_i = 1]}$$

Задача классификации

Множество всех возможных значений целевой функции **конечно**, каждое значение представляет собой отдельный класс, наша задача – угадывать правильный класс



Бинарная классификация



Многоклассовая классификация

Метрики

$y_1, y_2, \dots, y_n$ - ответы (целевые значения)

$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$- прогнозы (на тех же объектах)

Примеры метрик:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

Задача регрессии

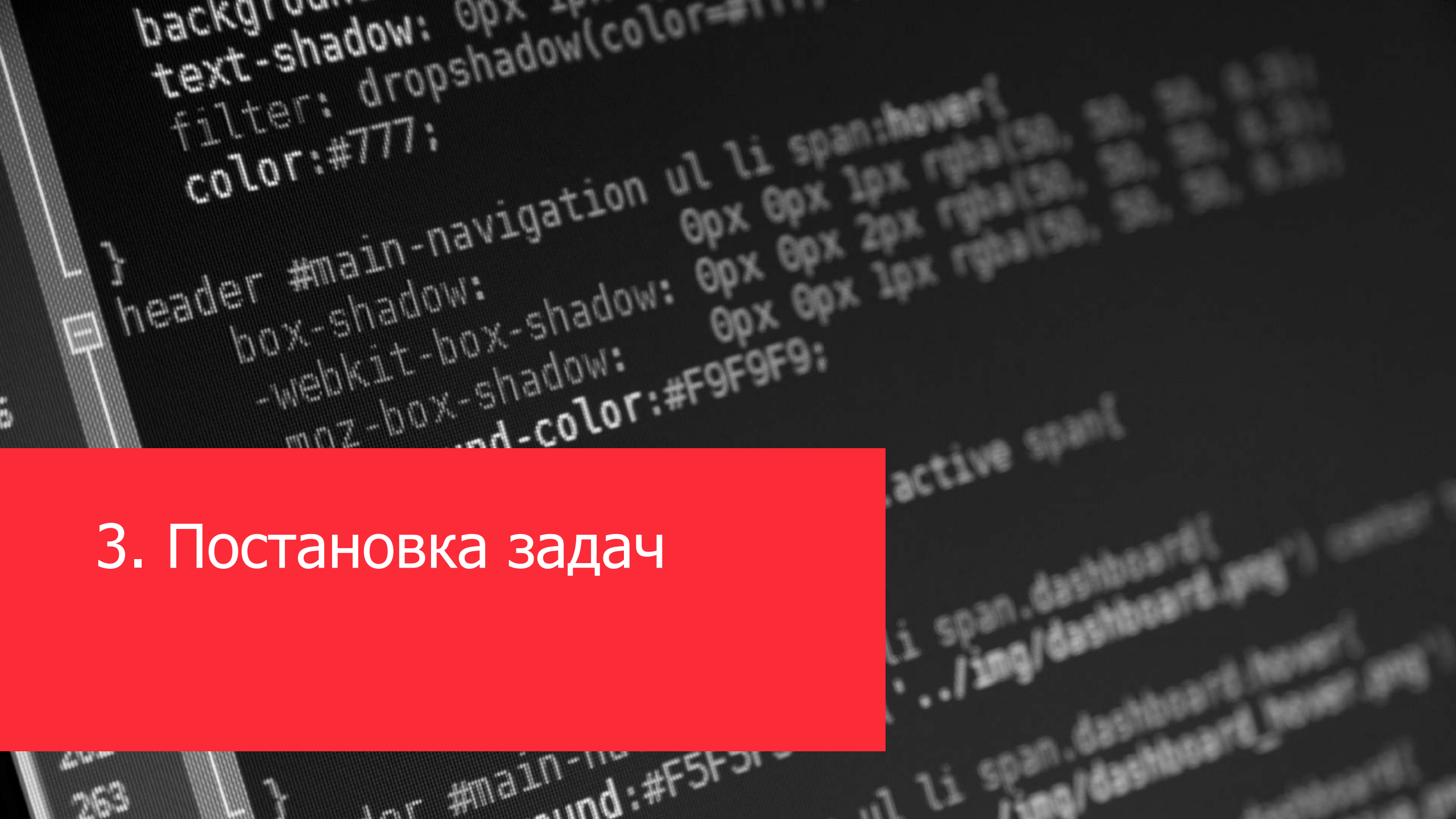
Значения целевой функции это числа, наша задача – делать прогнозы как можно ближе к правильному ответу

Вопросы

1. Что значит «спрогнозировать»?
2. Какие виды машинного обучения не попадают под такое описание?

Виды машинного обучения

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning
- Active learning
- …



3. Постановка задач

Постановка задачи

1. Что и зачем прогнозируем (что таргет, как используем прогноз)
2. Как оцениваем качество в оффлайне
3. Как оцениваем качество в онлайне

Отдельный важный вопрос:

Как оцениваем экономический эффект

Осмысленный выбор задачи

Очень важно не разрабатывать априори бесполезные модели.

Для этого:

1. задача должна быть решаемой
2. причем на том уровне качества, при котором экономика сходится

Пример: оптимизация колл-центра

Есть 100 сотрудников на телефонах и один контролирующий их супервайзер.

Вам поставлена задача оптимизировать их работу. Что делать?

Пример: оптимизация колл-центра

Вариант 1. Заменять людей ботами

Вариант 2. Заменять супервайзера

И там и там экономика не сошлась, что делать?

Оптимизация колл-центра: смотрим шире

Вариант 1. Заменять людей ботами

Вариант 2. Заменять супервайзера

И там и там экономика не сошлась, что делать?

Вариант 3. Прогнозируем загрузку колл-центра и выводим нужное число операторов.

Оптимизация колл-центра: смотрим шире

Вариант 1. Заменять людей ботами

Вариант 2. Заменять супервайзера

И там и там экономика не сошлась, что делать?

Вариант 3. Прогнозируем загрузку колл-центра и выводим нужное число операторов.

Вариант 4. Распределение звонков по операторам (увеличение производительности)

Правильный выбор метрики в оффлайне

Важно, чтобы метрика качества:

1. Не приводила при ее оптимизации к оценке не того, что нужно
2. Была связана с экономическими или другими важными показателями проекта

Правильный выбор метрики в оффлайне

Важно, чтобы метрика качества:

1. Не приводила при ее оптимизации к оценке не того, что нужно
2. Была связана с экономическими или другими важными показателями проекта

Кейс: оценка качества продукта

Задача: классифицировать производимый на производстве продукт на 3 класса - пригодный к использованию, непригодный, нельзя понять по имеющимся замерам

Кейс: оценка качества продукта

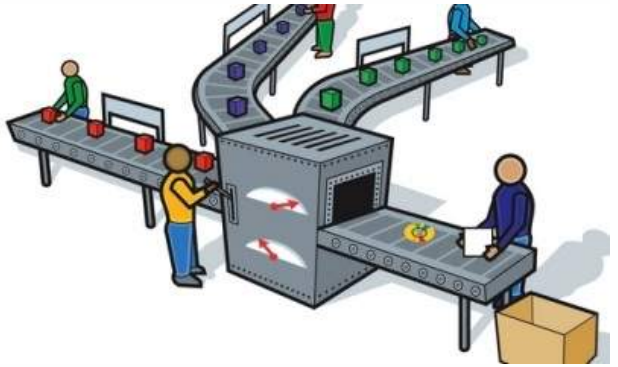
Задача: классифицировать производимый на производстве продукт на 3 класса - пригодный к использованию, непригодный, нельзя понять по имеющимся замерам

Выход алгоритма: три вероятности (p1, p2, p3)

Ответы в исторических данных: (1,0,0), (0, 1, 0), (0, 0, 1)

Выбор метрики качества:

Средний модуль покоординатной разности



Кейс: оценка качества продукта

Задача: классифицировать производимый на производстве продукт на 3 класса - пригодный к использованию, непригодный, нельзя понять по имеющимся замерам

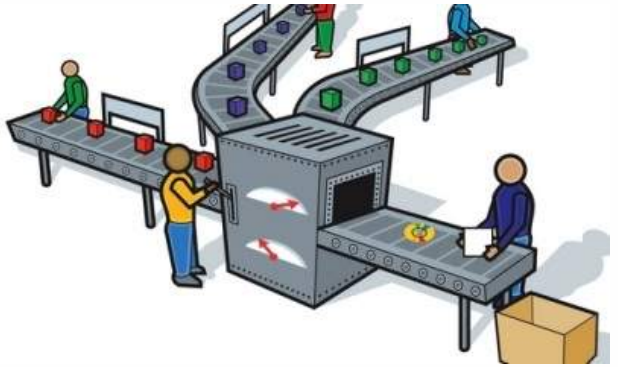
Выход алгоритма: три вероятности (p1, p2, p3)

Ответы в исторических данных: (1,0,0), (0, 1, 0), (0, 0, 1)

Выбор метрики качества:

Средний модуль покоординатной разности

???



Правильный выбор метрики в оффлайне

Важно, чтобы метрика качества:

1. Не приводила при ее оптимизации к оценке не того, что нужно
2. Была связана с экономическими или другими важными показателями проекта

Ответы на обращения клиентов

В отдел поддержки постоянно поступают запросы от клиентов организации

2016 год: 1000 обращений в день, успевают посмотреть все

2017 год: 5000 обращений в день, успевают посмотреть 2000 (фильтр по ключевым словам)

2018 год: как выжить с 20000 обращений?

Экономика проекта

Где здесь деньги:

По простым фильтрам за день выделяется 2000 важных обращений, из них действительно срочных – 800

Экономика проекта

Где здесь деньги:

По простым фильтрам за день выделяется 2000 важных обращений, из них действительно срочных – 800

Итого – поддержка работает на 40%

Экономика проекта

Где здесь деньги:

По простым фильтрам за день выделяется 2000 важных обращений, из них действительно срочных – 800

Итого – поддержка работает на 40%

Как обеспечить ту же эффективность на большем объеме обращений без расширения штата? (Либо повысить эффективность)

Какую задачу решаем

Вывод: нужно классифицировать обращения на более срочные и менее срочные лучше, чем фильтр по ключевым словам.

Как оцениваем

Вывод: нужно классифицировать обращения на более срочные и менее срочные лучше, чем фильтр по ключевым словам.

Пример: если в топ-2000 обращений по ML будет 1000 срочных, то тем же отделом поддержки обработаем на 25% больше обращений.

Как оцениваем

Вывод: нужно классифицировать обращения на более срочные и менее срочные лучше, чем фильтр по ключевым словам.

Пример: если в топ-2000 обращений по ML будет 1000 срочных, то тем же отделом поддержки обработаем на 25% больше обращений.

Сэкономили на расширении 25% затрат на сотрудников отдела (грубо)

Оценка качества в онлайне

1. Должен быть заранее придуманный дизайн A/B теста
2. Должен быть способ оценки статистической значимости результата
3. Должна быть одна главная метрика, по которой принимается решение: действительно важная для проекта, не слишком шумная, такая, которую можем статистически значимо измерить
4. На другие метрики смотрим «для проверки» – чтобы заметить, если что-то совсем сломалось

Заранее утвержденный дизайн A/B

Нельзя сначала получить результаты, а потом искать срез/метрику/разбиение на группы/фильтрации групп при которых будет результат от алгоритма. Всегда найдете.

Нельзя продлевать A/B до получения результата (если только не используете методы последовательной проверки гипотез), нужно сразу определиться с длительностью теста

Оценка статистической значимости

Предположения критерия должны выполняться (пример: если считаем распределение нормальным – оно должно быть таким)

Альтернатива против которой рассматриваем гипотезу тоже должна соответствовать вашей ситуации (пример: если считаем, что распределения в группах отличаются только сдвигом, то так и должно быть)

Выбор основной метрики

Пример: рекомендации товаров в e-commerce

Главная по смыслу метрика – доход в группе

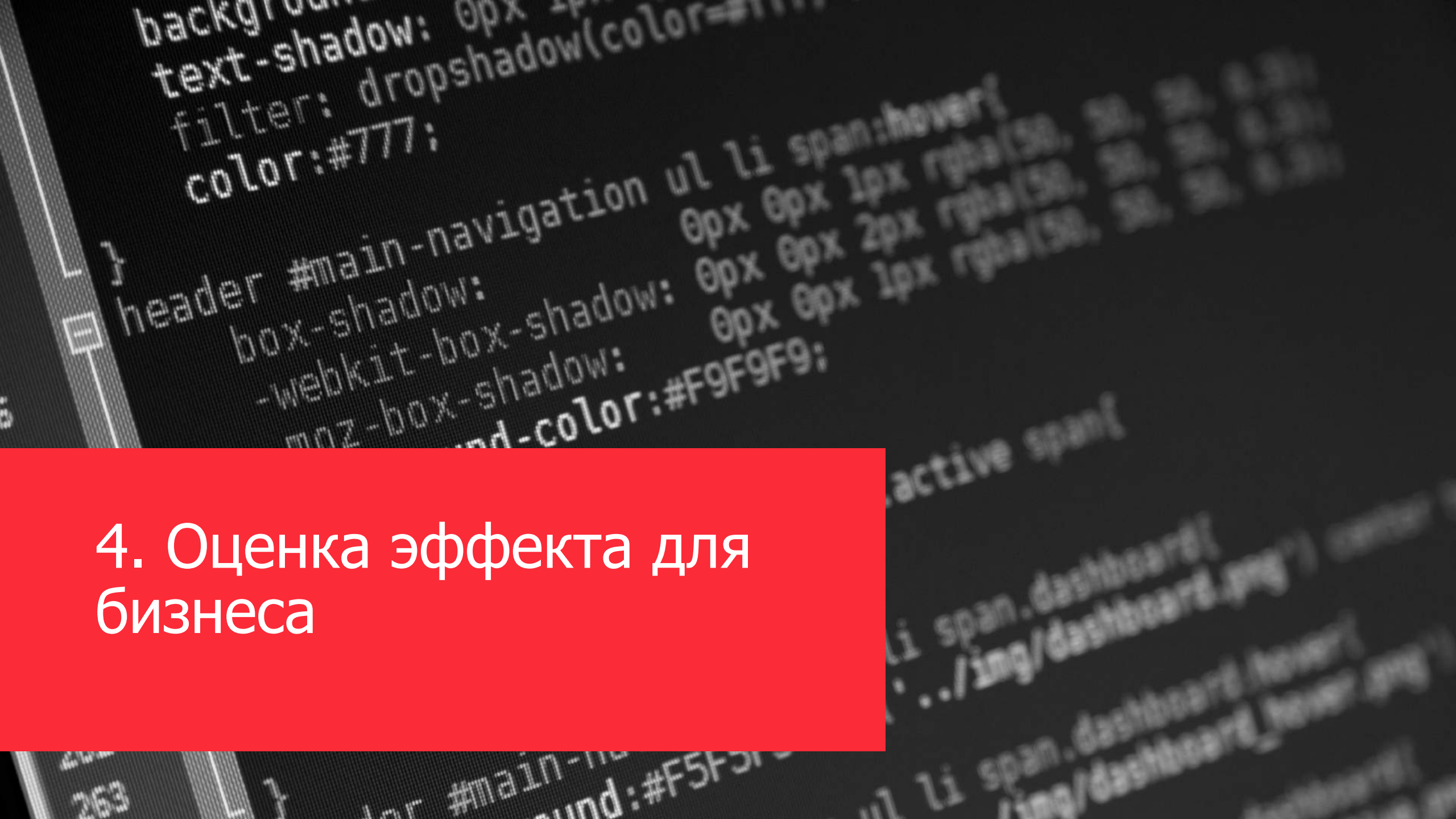
Но «деньги незначимы»

Приходится использовать прокси-метрики – например, конверсию в покупки

Мониторинг других метрик

Если увеличим конверсию, просадив средний чек, принципиально ничего не изменится.

Поэтому средний чек тоже нужно мониторить, даже если решение договорились принимать по конверсии



4. Оценка эффекта для бизнеса

Оценка экономического эффекта

1. Иногда можно оценить напрямую
2. Иногда эффект не очевиден, но можно прикинуть косвенно
3. Также можно перевести в неденежный, но тоже важный показатель

Пример оценки эффекта напрямую

Прогнозируем отток пользователей из приложения, чтобы тех, кто хочет от нас уйти, удержать скидкой на 10% на год

Пример оценки эффекта напрямую

Прогнозируем отток пользователей из приложения, чтобы тех, кто хочет от нас уйти, удержать скидкой на 10% на год

В горизонте 1 года:

ARPU – зарабатываем в среднем с пользователя

0.1 × ARPU – тратим на удержание пользователя

N – количество пользователей, которым даем ссылку

CF – доля тех, что уйдет в отток без удержания

0 – доля тех, кто уйдет с удержанием :) (рассмотрим простой случай)

Пример оценки эффекта напрямую

Прогнозируем отток пользователей из приложения, чтобы тех, кто хочет от нас уйти, удержать скидкой на 10% на год

В горизонте 1 года:

Сэкономленные деньги = $ARPU \times N \times CF - 0.1 \times ARPU \times N$

Пример оценки эффекта напрямую

Прогнозируем отток пользователей из приложения, чтобы тех, кто хочет от нас уйти, удержать скидкой на 10% на год

В горизонте 1 года:

Сэкономленные деньги = $ARPU \times N \times CF - 0.1 \times ARPU \times N$

Т.е. если CF меньше 0.1 даже бесплатная разработка модели не окупится, а с учетом стоимости разработки модели, и с учетом того, что не всех скидка удержит – все сложнее

Как метрики связаны с экономикой

В горизонте 1 года:

Сэкономленные деньги = $ARPU \times N \times CF - 0.1 \times ARPU \times N$

Чем больше CF – тем больше денег экономим.

Что такое CF? Это доля действительно собиравшихся в отток из удерживаемых N пользователей.

Как метрики связаны с экономикой

В горизонте 1 года:

Сэкономленные деньги = $ARPU \times N \times CF - 0.1 \times ARPU \times N$

Чем больше CF – тем больше денег экономим.

Что такое CF? Это доля действительно собиравшихся в отток из удерживаемых N пользователей.

Вывод 1: удерживаем N пользователей с максимальной вероятностью оттока

Как метрики связаны с экономикой

В горизонте 1 года:

Сэкономленные деньги = $ARPU \times N \times CF - 0.1 \times ARPU \times N$

Чем больше CF – тем больше денег экономим.

Что такое CF? Это доля действительно собиравшихся в отток из удерживаемых N пользователей.

Вывод 1: удерживаем N пользователей с максимальной вероятностью оттока

Вывод 2: уместная метрика - доля класса 1 среди топ N по вероятности. Такая метрика есть, называется Recall@N.

Как метрики связаны с экономикой

В горизонте 1 года:

Сэкономленные деньги = $ARPU \times N \times CF - 0.1 \times ARPU \times N$

Чем больше CF – тем больше денег экономим.

Что такое CF? Это доля действительно собиравшихся в отток из удерживаемых N пользователей.

Вывод 1: удерживаем N пользователей с максимальной вероятностью оттока

Вывод 2: уместная метрика - доля класса 1 среди топ N по вероятности. Такая метрика есть, называется Recall@N. А могли же ROC-AUC начать мерять! :)

Как метрики связаны с экономикой

В горизонте 1 года:

Сэкономленные деньги = ARPU × N × CF – 0.1 × ARPU × N

Чем больше CF – тем больше денег экономим.

Что такое CF? Это доля действительно собиравшихся в отток из удерживаемых N пользователей.

Вывод 1: удерживаем N пользователей с максимальной вероятностью оттока

Вывод 2: уместная метрика - доля класса 1 среди топ N по вероятности. Такая метрика есть, называется Recall@N. А могли же ROC-AUC начать мерять! :)

Примечание: на самом деле в прогнозе оттока чаще смотрят на очень близкую к Recall@N метрику – Lift 10 = Recall@10% / Recall @100%

Trade-off при оценке экономики

1. Делаем очень подробную экономическую модель, учитывающую множество факторов и использующую множество волшебных чисел («вероятность удержания» и т.п.) которые мы оценили очень примерно, и они сами вносят погрешность в оценку

Trade-off при оценке экономики

1. Делаем очень подробную экономическую модель, учитывающую множество факторов и использующую множество волшебных чисел («вероятность удержания» и т.п.) которые мы оценили очень примерно, и они сами вносят погрешность в оценку
2. Делаем очень простую экономическую модель, в которой много допущений, но мало волшебных констант

Trade-off при оценке экономики

1. Делаем очень подробную экономическую модель, учитывающую множество факторов и использующую множество волшебных чисел («вероятность удержания» и т.п.) которые мы оценили очень примерно, и они сами вносят погрешность в оценку
2. Делаем очень простую экономическую модель, в которой много допущений, но мало волшебных констант

Оптимальный вариант, конечно, посередине

Косвенная оценка эффекта

Внедряем в мобильное uber-like приложение подсказки для заказа

Сами подсказки денег вроде бы не приносят, только делают приложение удобнее для пользователя

Косвенная оценка эффекта

Внедряем в мобильное uber-like приложение подсказки для заказа

Сами подсказки денег вроде бы не приносят, только делают приложение удобнее для пользователя

Допустим, мы уже сделали проект: как грубо оценить влияние подсказок на деньги?

Косвенная оценка эффекта

Внедряем в мобильное uber-like приложение подсказки для заказа

Сами подсказки денег вроде бы не приносят, только делают приложение удобнее для пользователя

Допустим, мы уже сделали проект: как грубо оценить влияние подсказок на деньги?

Измеряем в A/B статзначимое повышение в конверсии, умножаем на оборот – это и есть очень грубая оценка (как правило, получаются большие числа :)

Косвенная оценка эффекта

Допустим, мы уже сделали проект: как грубо оценить влияние подсказок на деньги?

Измеряем в A/B статзначимое повышение в конверсии, умножаем на оборот – это и есть очень грубая оценка (как правило, получаются большие числа :)

А как оценить это, если еще не сделали проект?

Косвенная оценка эффекта

А как оценить это, если еще не сделали проект?

Вариант 1. Ухудшаем подсказки (если уже есть), измеряем понижение конверсии на небольшой группе, понимаем, как качество подсказок влияет на конверсию, пересчитываем.

Косвенная оценка эффекта

А как оценить это, если еще не сделали проект?

Вариант 1. Ухудшаем подсказки (если уже есть), измеряем понижение конверсии на небольшой группе, понимаем, как качество подсказок влияет на конверсию, пересчитываем.

Вариант 2. Подсказки сокращают время заказа. Ухудшим его искусственно и посмотрим на его влияние на конверсию.

Оценка других показателей

Продолжая историю с подсказками в приложении: если они уменьшают время, которое тратится на заказ, то измеряем этот эффект.

Оценка других показателей

Продолжая историю с подсказками в приложении: если они уменьшают время, которое тратится на заказ, то измеряем этот эффект.

Умножаем эффект на размер пользовательской базы и среднее число заказов в день, делим на среднюю продолжительность жизни, публикуем пресс-релиз про то как машинное обучение в подсказках сохраняет N человеческих жизней в день :)

Оценка других показателей

Продолжая историю с подсказками в приложении: если они уменьшают время, которое тратится на заказ, то измеряем этот эффект.

Умножаем эффект на размер пользовательской базы и среднее число заказов в день, делим на среднюю продолжительность жизни, публикуем пресс-релиз про то как машинное обучение в подсказках сохраняет N человеческих жизней в день :)

Но, кроме шуток, время пользователя – вполне осязаемый и важный показатель. Ориентироваться на него вполне нормально.

Почему не всегда оценивают экономику

Пример: задача кредитного скоринга

Экономику оценивают, но, увы, часто уже после проекта

Почему: скоринг делается для нескольких кредитных продуктов, и у каждого своя логика выдачи, свои отсечки по скору и свой учет размера кредита (и делать это могут разные подразделения)

У каждого продукта этот процесс может меняться относительно независимо

Поэтому экономическая модель получается слишком сложной и ее нужно будет регулярно обновлять

Почему не всегда оценивают экономику

В итоге, несмотря на то, что в скоринге тоже важен Recall@N, как правило прогноз оценивают по метрике Gini = 2 × ROC-AUC – 1

Т.к. метрика уже устоялась – перейти от нее к метрикам по топам организационно почти невозможно, но можно измерять параллельно

Также можно грубо оценивать как рост Gini на 1-2% влияет на экономику.
Вопрос: чем опасны такие оценки?

Специфика IT-компаний

1. В подразделениях, существующих больше нескольких лет, постановки задач уже зафиксированы и основная работа – улучшение существующих моделей и разработка новых, более точных.
2. В новых подразделениях много новых задач, и там вопрос правильной постановки более актуален.

Обсудили сегодня

1. Организационные моменты
2. Задача машинного обучения
3. Постановка задач в терминах машинного обучения
4. Оценка эффекта для бизнеса

Обсудили сегодня

1. Организационные моменты
2. Задача машинного обучения
3. Постановка задач в терминах машинного обучения
4. Оценка эффекта для бизнеса

Не расходитесь, после перерыва будет семинар :)

Что будет через неделю

Линейные модели в классификации и регрессии

Как работают и где используются линейные модели. Задачи с разреженными признаками. Особенности использования линейных моделей: онлайн-обучение, комбинирование методов оптимизации. Повышение устойчивости линейных моделей.

Hashing trick и Vowpal Wabbit.