

# Машинное обучение

Лекция 7. Постановка задач



# История о постановке задач

На входе:

- Туropратор хочет персонализировать рассылки своим клиентам
- Данных мало
- Заказчику интересно увеличить конверсию

# История о постановке задач

## На входе:

- Туropратор хочет персонализировать рассылки своим клиентам
- Данных мало
- Заказчику интересно увеличить конверсию

## На выходе:

- Не стали надеяться на какое-то умное обучение
- Попытались кластеризовать и выделить топ рекомендаций по кластерам
- Отдали заказчику в надежде, что в любом случае лучше, чем всем рассылать одно и то же
- Про А/В тесты вообще не слышали

# Обсуждаем сегодня: постановка задач

- I. Что прогнозировать и как использовать модель
- II. Как оценивать качество (онлайн и оффлайн)
- III. Выборка, признаки и модель: насколько это важно
- IV. Примеры задач, где не всё просто

# I. Что прогнозировать и как использовать модель

## Таргет и использование модели

1. Какую величину надо прогнозировать
2. Зачем это нам: как мы будем ее использовать
3. Какой эффект мы можем получить
4. Нужно ли учесть какие-то дополнительные сложности

Какую величину надо прогнозировать

**Пример:**

Пользователи ставят рейтинги товарам.

**Предложение:**

Давайте будем прогнозировать, какой рейтинг поставит пользователь товару

*Чем это плохо?*

## Пример: рекомендации

$$P(\text{покупка} \mid \text{показ}) = P(\text{покупка} \mid \text{клик}) P(\text{клик} \mid \text{показ})$$



мало данных

чуть больше данных

# Использование прогноза

**Пример:**

Туроператор хотел получить модель, прогнозирующую вероятность, что клиент купит тур.

**Вопрос, поставивший их «в ступор»:**

Что будете с ней делать?

# Использование прогноза: пример про UX

Сопутствующие товары	Похожие товары		
<b>Товар 1</b>	<b>Товар 2</b>	<b>Товар 3</b>	<b>Товар 4</b>

# Какой эффект можем получить

Пример: давайте разработаем программу повышения конверсии новичков в наших постоянных клиентов?

# Какой эффект можем получить

Пример: давайте разработаем программу повышения конверсии новичков в наших постоянных клиентов?

## Довод 1:

Новичков 1%, значит даже если мы сделаем на 5-10% больше заказов от новичков, это даст всего 0,05-0,010% дополнительных заказов

# Какой эффект можем получить

Пример: давайте разработаем программу повышения конверсии новичков в наших постоянных клиентов?

## Довод 1:

Новичков 1%, значит даже если мы сделаем на 5-10% больше заказов от новичков, это даст всего 0,05-0,010% дополнительных заказов

## Довод 2:

Все новые клиенты сначала становятся «новичками», вопрос в долгосрочном эффекте

# Какой эффект можем получить

Другой пример:

Будем допродавать услуги пользователям (те, которые наиболее вероятно купят после рекомендации).

# Какой эффект можем получить

Другой пример:

Будем допродавать услуги пользователям (те, которые наиболее вероятно купят после рекомендации).

Оценка возможного эффекта:

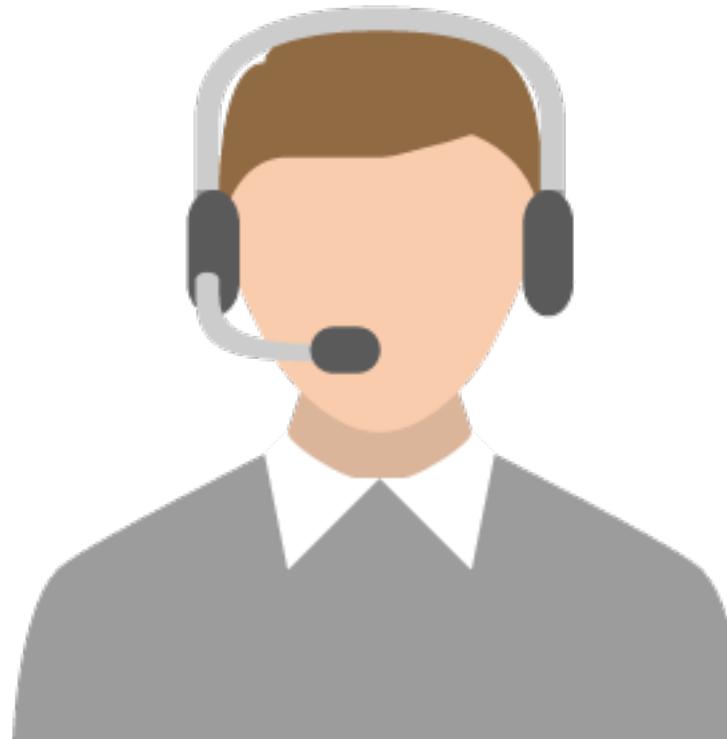
- Допустим, в 1-2% случаев раз в месяц пользователи будут соглашаться на допродажу
- Средняя стоимость услуги 1000 рублей, пользователей 100000
- Значит можем получить +1-2 млн. рублей в месяц, откуда следуют оценки на бюджет проекта и его приоритетность

## Дополнительные сложности

Пример 1: Разброс в таргете.

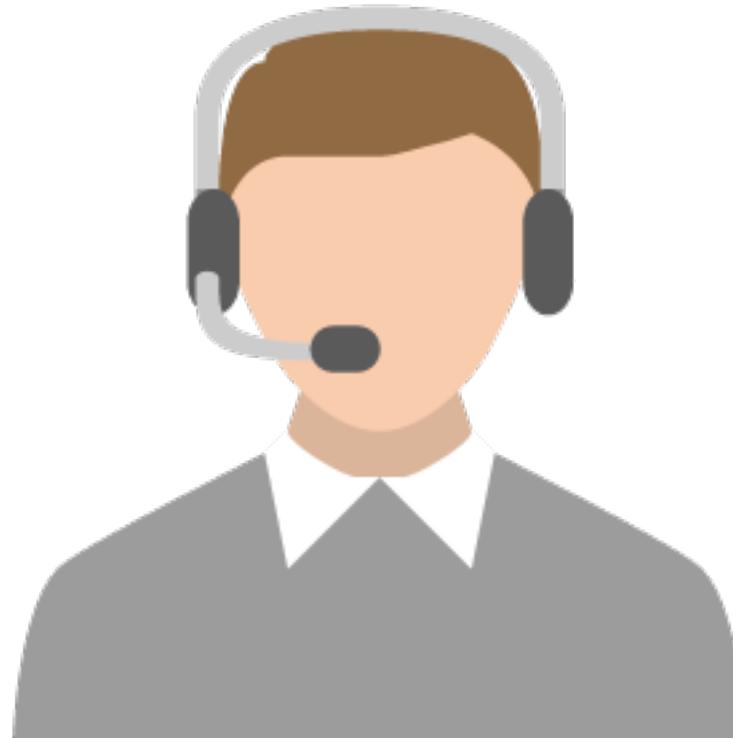
Очень часто встречается, когда прогнозируется субъективная оценка людей. Например, тональность отзыва о компании.

## Пример: обращения в службу поддержки



- Каждый день служба поддержки получает тысячи текстовых обращений
- Некоторые из них – о ситуациях, которые могут нанести большой ущерб репутации компании
- Такие кейсы обрабатываются в срочном порядке специальной командой

## Пример: обращения в службу поддержки



- «Опасные» обращения фильтруются по набору ключевых слов
- Отфильтровывается 5% всех обращений
- Среди них всего 30% - действительно опасные
- Команда поддержки в ближайшее время не расширяется, но нужно обрабатывать больше «опасных» обращений

## Решение 1

1. Построить модель, по тексту обращения определяющую вероятность того, что оно «срочное»
2. Заменить/дополнить фильтр такой моделью

## Проблема с таргетом

Проблема: качество работы модели оказалось незначительно лучше, чем у фильтра

В итоге выяснилось, что совпадение в разметке от разных сотрудников – порядка 50%

## Решение 2

1. Построить модель, по тексту обращения определяющую вероятность того, что оно «срочное»
2. Ранжировать обращения по вероятности, оцененной моделью
3. Дообучать модель при использовании

## Дополнительные сложности

**Пример 2:** ошибка в таргете может «прощаться»

Допустим, мы хотим угадать, куда поедет пользователь, и угадываем с точностью до соседнего дома. Будет ли это правильным ответом?

В подобных ситуациях можно записывать в обучающую выборку как «правильный» не только идеальный ответ, но и другие подходящие.

## Дополнительные сложности

Пример 3: Зависимость не существует.

Можно смотреть на эту ситуацию как на разброс в таргете, с той лишь разницей, что такой таргет не получится собирать более точно.

## Дополнительные сложности

**Пример 3:** Зависимость не существует.

Можно смотреть на эту ситуацию как на разброс в таргете, с той лишь разницей, что такой таргет не получится собирать более точно.

Пример: вы хотите прогнозировать результат работы какого-то другого алгоритма (например, алгоритма ваших конкурентов). Но алгоритм меняется каждый день. В этом случае исторические данные помогут вам восстановить «зависимость» только очень приближенно

## Дополнительные сложности

Пример 4:

Произошли сильные изменения в зависимости, делающие исторические данные бесполезными

## II. Как оценивать качество

## Оценка качества

1. Какую величину надо оптимизировать «в жизни»
2. Что надо оптимизировать на исторических данных
3. Что и как надо измерять при использовании модели
4. Trade-off между мониторингом качества и максимальным эффектом
5. Что делать, если оценка качества пошла не по плану

# Какую величину оптимизируем «в жизни»

Пример:

Удерживаем пользователей – оптимизируем стоимость для нас денег, заработанных на пользователе.

Вариант 1:

Если получаем с пользователя  $X$  рублей за все его «время жизни» ( $X$  обычно называют LTV), а добиться того, чтобы он был с нами нам стоит  $Y$  рублей, то с каждого потраченного рубля мы получили в  $X/Y$  раз больше и логично это отношение максимизировать.

# Какую величину оптимизируем «в жизни»

Пример:

Удерживаем пользователей – оптимизируем стоимость для нас денег, заработанных на пользователе.

Вариант 2:

Если получаем с пользователя  $X$  рублей в месяц, а добиться того, чтобы он был с нами нам стоит  $Y$  рублей в месяц, то с каждого потраченного рубля мы получаем в  $X/Y$  раз больше и логично это отношение максимизировать.

# Какую величину оптимизируем «в жизни»

## Пример:

Удерживаем пользователей – оптимизируем стоимость для нас денег, заработанных на пользователе.

## Особенности:

- Может быть *не выгодно* удерживать пользователей
- Можно упереться в ширину канала привлечения

# Что оптимизируем на исторических данных

Пример:

Удерживаем пользователей

Вопрос 1: Кого нужно удерживать?

Вопрос 2: Какая метрика нам важна?

## Вернемся к экономике

- Пусть мы пытаемся удержать  $N$  пользователей, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого  $C$  денег
- $r$  – доля настоящих отточников среди удерживаемых
- $N_p$  – удержаных пользователей, если удерживаем со 100% успехом
- $N_{pr}$  – удержаных пользователей, если удержание успешно с вероятностью  $r$

## Вернемся к экономике

- Пусть мы пытаемся удержать  $N$  пользователей, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого  $C$  денег
- $r$  – доля настоящих отточников среди удерживаемых
- $Npr$  – удержаных пользователей, если удержание успешно с вероятностью  $r$

**Вывод:**

Наша задача – максимизировать  $r$

Экономический эффект:  $ARPU * N * p * r - C * N$

## Вернемся к экономике

- Пусть мы удерживаем  $N$  пользователей, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого  $C$  денег
- $r$  – доля настоящих отточников среди удерживаемых
- $Npr$  – удержанных пользователей, если удержание успешно с вероятностью  $r$

Precision@N

**Вывод:**

Наша задача – максимизировать  $r$

Экономический эффект:  $ARPU * N * p * r - C * N$

## Выбор метрики в задаче удержания

Несмотря на наши выкладки, в задаче прогнозирования оттока традиционно смотрят не на  $\text{precision}@N$  из-за особенности данных.  
Но об этом – в лекции про предиктивную аналитику.

# Онлайновая оценка качества

Допустим, на исторических данных качество алгоритма высокое, а будет ли оно высоким в реальности?

Идеи:

1. А/В тест
2. Оценка статзначимости результата

## A/B тест

1. Случайным образом делим клиентов на равные группы
2. Измеряем целевые метрики (например, доход с каждой группы клиентов) в каждой группе за длительный период времени
3. Получаем какое-то число для каждой группы
4. Что дальше?

# Статистическая значимость: пример



# Статистическая значимость: пример



Одна кривая отличается от других на 10%  
Но разбиение на самом деле – случайное

# Измерения в онлайне на примере рекомендаций

- Доход в группе
- Доход с пользовательской сессии
- Средняя стоимость купленного товара
- Средний чек
- Конверсия в покупку
- Клики
- Различные модели атрибуции: last click, first click

# Напоминание: бутстреп

Выборка:

№	X
1	3.4
2	2.9
3	3.7
N	3.1

# Бутстреп

Выборка:

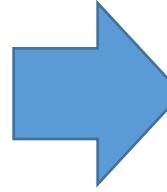
№	X
1	3.4
2	2.9
3	3.7
N	3.1

$$\mathbb{E} X = 3.3 \pm ?$$

# Бутстреп

Выборка:

№	X
1	3.4
2	2.9
3	3.7
N	3.1



№	X
3	3.7
1	3.4
2	2.9
2	3.1

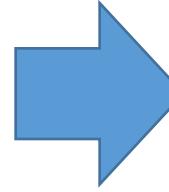
Генерируем новую  
(искусственную) выборку,  
отбрав в нее объекты из  
исходной выборки по  
схеме выбора с  
возвращением

$$\mathbb{E} X = 3.3 \pm ?$$

# Бутстреп

Выборка:

№	X
1	3.4
2	2.9
3	3.7
N	3.1



№	X
3	3.7
1	3.4
2	2.9
2	3.1

№	X
1	3.4
3	3.7
2	2.9
M	3.0

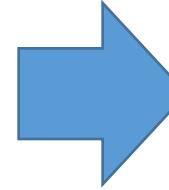
Продолжаем  
генерировать  
такие выборки

$$\mathbb{E} X = 3.3 \pm ?$$

# Бутстреп

Выборка:

№	X
1	3.4
2	2.9
3	3.7
N	3.1



№	X
3	3.7
1	3.4
2	2.9
2	2.9

№	X
1	3.4
3	3.7
2	2.9
M	3.0

...

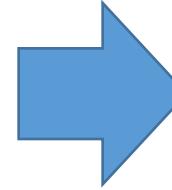
№	X
3	3.7
2	2.9
1	3.4
1	3.4

$$\mathbb{E} X = 3.3 \pm ?$$

# Бутстреп

Выборка:

№	X
1	3.4
2	2.9
3	3.7
N	3.1



№	X
3	3.7
1	3.4
2	2.9
2	2.9

№	X
1	3.4
3	3.7
2	2.9
M	3.0

...

№	X
3	3.7
2	2.9
1	3.4
1	3.4

$$\mathbb{E} X = 3.31 \pm ?$$

$$\mathbb{E} X = 3.25$$

$$\mathbb{E} X = 3.27$$

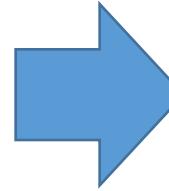
...

$$\mathbb{E} X = 3.39$$

# Бутстреп

Выборка:

№	X
1	3.4
2	2.9
3	3.7
N	3.1



№	X
3	3.7
1	3.4
2	2.9
2	2.9

№	X
1	3.4
3	3.7
2	2.9
M	3.0

...

№	X
3	3.7
2	2.9
1	3.4
1	3.4

$$\mathbb{E} X = 3.31 \pm ?$$

$$\mathbb{E} X = 3.25$$

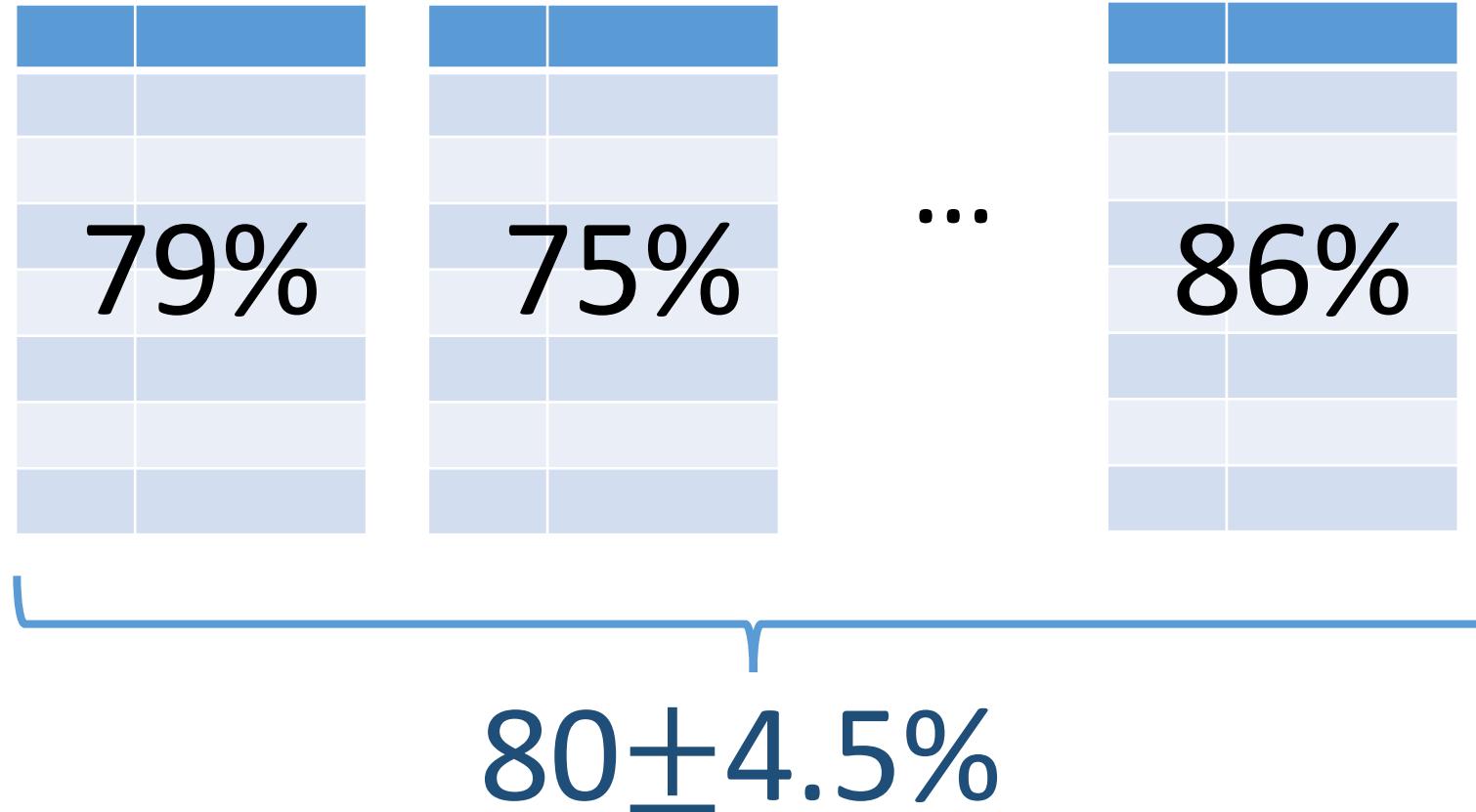
$$\mathbb{E} X = 3.27$$

...

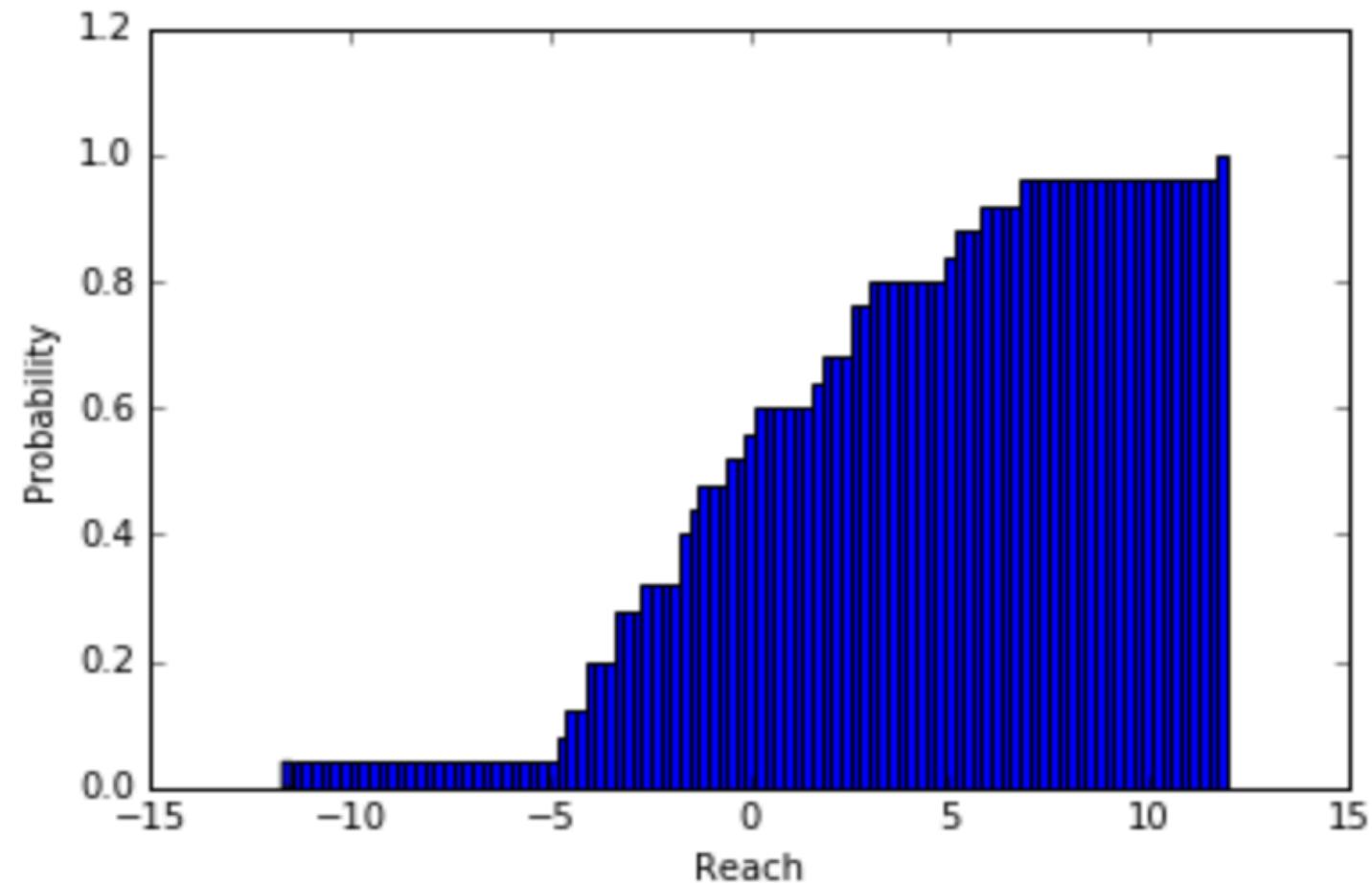
$$\mathbb{E} X = 3.39$$

$$\mathbb{E} X = 3.32 \pm 0.06$$

# Бутстреп



# Эмпирическое распределение



### **III. Выборка, признаки и модель**

## Пример: оценка вероятности покупки

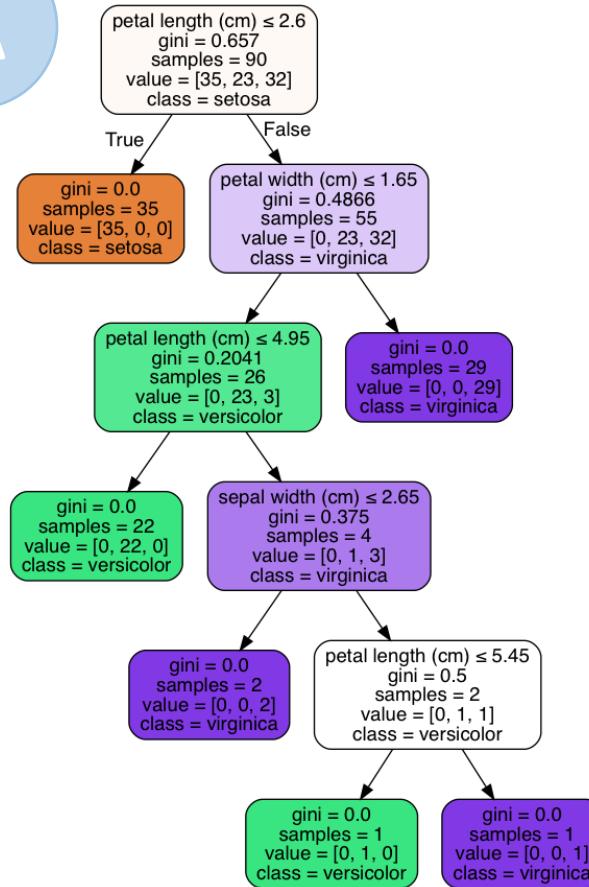
- Объекты: тройки (пользователь, товар, момент времени)
- Классы: 1 - товар будет куплен, 0 – товар не будет куплен
- Признаки: параметры пользователя, товара, момента времени и их «взаимодействие»

## Генерация негативных примеров

- Добавить к каждому позитивному примеру весь каталог как негативный (не реально)
- Случайные с равномерным распределением
- Случайные, с вероятностями, пропорциональными популярности объекта
- Самые популярные примеры
- Те объекты, которые рекомендовал бы какой-то алгоритм, но они не были куплены

# Выбор модели

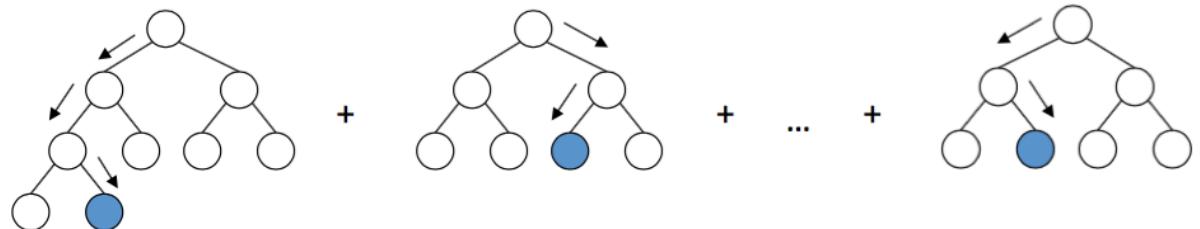
A



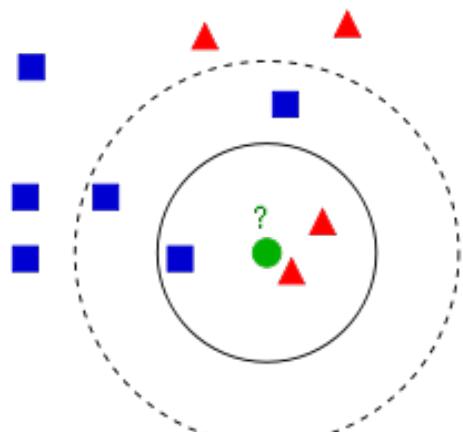
B

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

C



D

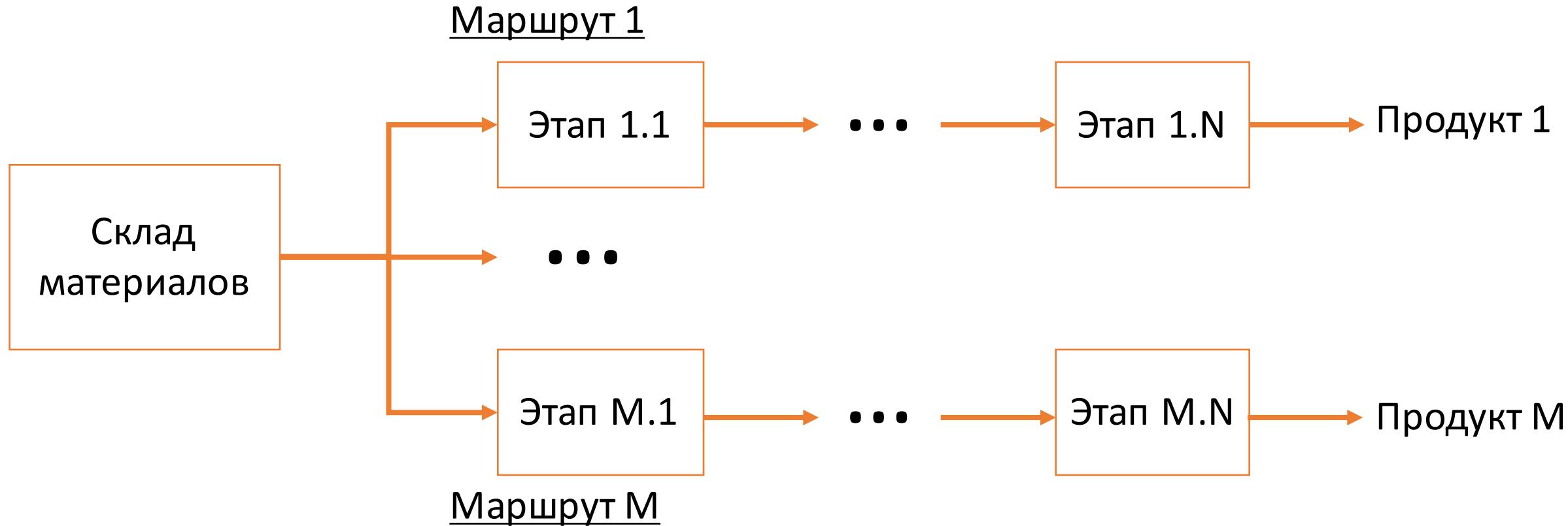


## IV. Примеры задач

# Прогнозирование дефектов при производстве



# Маршруты производства



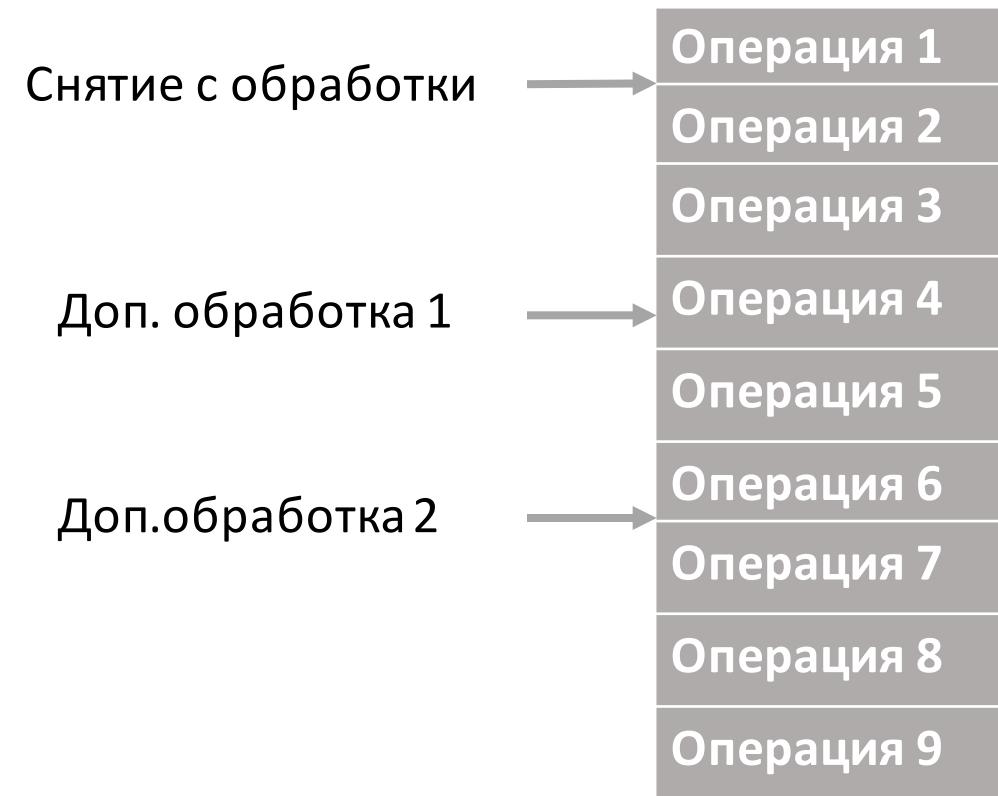
## Дополнительная обработка

Между некоторыми обязательными этапами обработки часть продукции может отправляться на доработку (чтобы уменьшить вероятность брака)

# Опишем процесс

	П1	П2	П3	П4	П5	П6	П7	П8
Операция 1	+	+	+	+	+	+	+	+
Операция 2	+	+	+	+	+	+	+	+
Операция 3	+	+			+	+		+
Операция 4	+	+	+	+	+	+	+	+
Операция 5	+	+			+	+		+
Операция 6				+				
Операция 7						+		
Операция 8						+		
Операция 9						+	+	

# Место дополнительной обработки в процессе



# Где здесь машинное обучение

- Учимся на этапах перед возможной доп.обработкой прогнозировать дефект в продукции
- Если дефект – бинарная величина, прогнозируем его вероятность (решаем задачу классификации), если вещественная (например, масса непригодного продукта), то прогнозируем матожидание (задача регрессии)
- Отправляем на доп.обработку продукцию, которой она больше всего нужна

## Экономика процесса

- Количество потраченных денег = потери из-за производства непригодного продукта (брата) + потери на доп. обработки и снятие с производства
- Доп. обработки уменьшают первое слагаемое и увеличивают второе
- Потери из-за брака = сумма потерь на каждом произведенном продукте
- Значит, зная матожидание потерь на каждом продукте – понимаем, какие

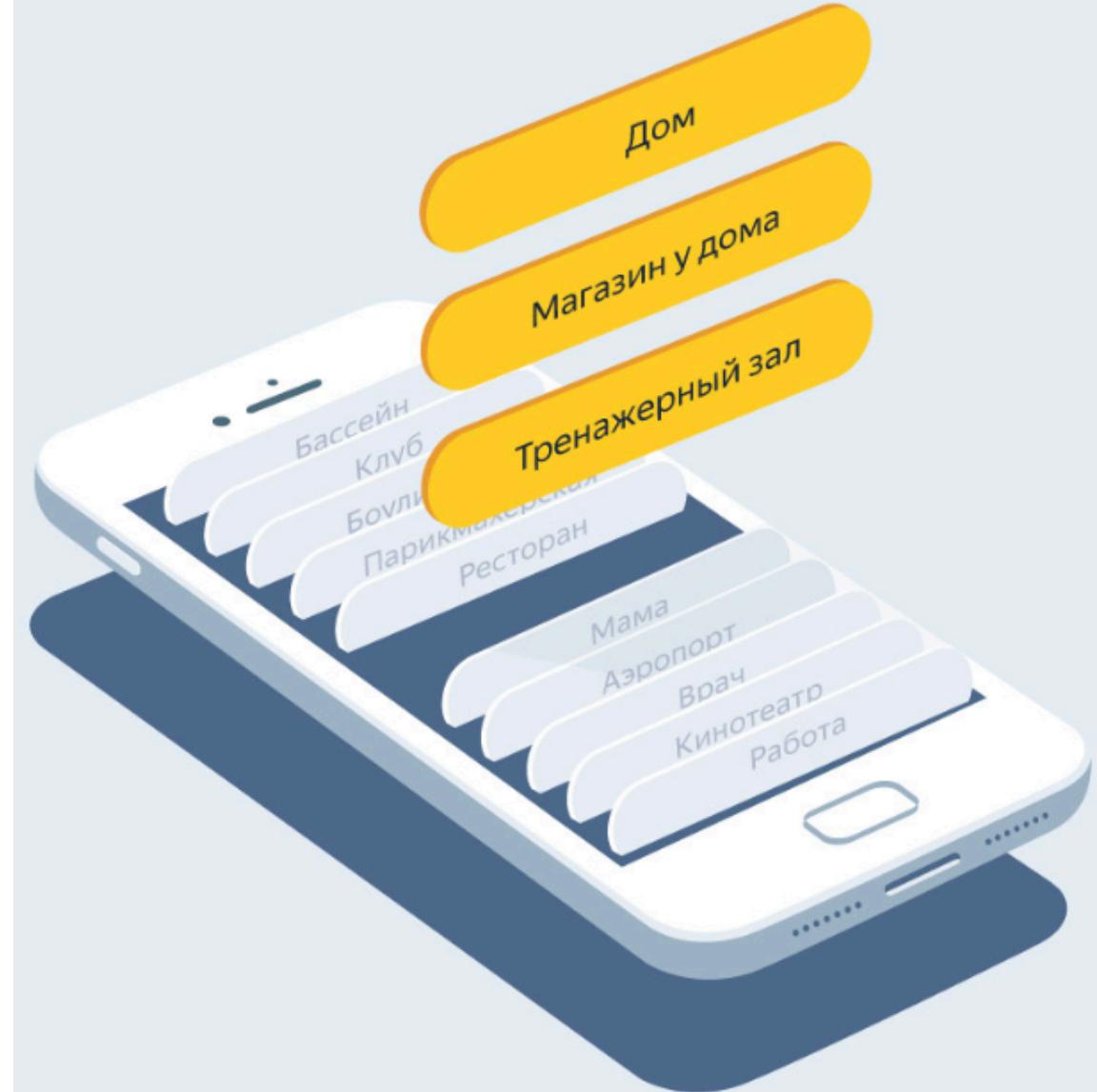
## Что еще можно обсудить

- Что включать в потери из-за снятия с производства
- Оценка того, насколько дополнительная обработка поможет устранить дефект
- Смещенность выборок

# Что вызовет больше всего проблем

- Нехватка данных для обучения
- Нехватка данных для теста со статзначимым результатом
- Невозможность в точности реализовать рекомендации модели (работники на местах все равно будут пытаться делать «по-своему»)

# Прогнозирование точки Б в Такси\*



\* Кейс публичный, прочитать подробно можно здесь:  
<https://habrahabr.ru/company/yandex/blog/337328/>

# Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В

## Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В
2. Задача регрессии: объекты – пользователи, целевое значение – координаты точки В

## Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В
2. Задача регрессии: объекты – пользователи, целевое значение – координаты точки В
3. Задача классификации: объекты - пары (пользователи, момент времени), классы – адреса мест, куда может поехать

## Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В
2. Задача регрессии: объекты – пользователи, целевое значение – координаты точки В
3. Задача классификации: объекты - пары (пользователи, момент времени), классы – адреса мест, куда может поехать
4. Задача классификации: объекты - тройки (пользователь, место, момент времени, классы – 1 и 0 (поехал или нет), 1 при точном попадании в адрес или в окрестность 20 метров

# Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)

# Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)
2. ROC-AUC

# Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)
2. ROC-AUC
3. Precision@1, Precision@2

# Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)
2. ROC-AUC
3. Precision@1, Precision@2
4. Recall@1, Recall@2

# Признаки

## ПРИЗНАКИ

## ПРИМЕРЫ

Данные по заказам пользователя

- Средние параметры поездки (время, расстояние)
- Частота заказов

Признаки текущей точки старта и предполагаемой точки назначения

- Количество поездок из или в эту точку
- Различные условные вероятности поездки

Признаки всего набора кандидатов, генерируемого по данной истории пользователя

- Общее число кандидатов
- Свойства списка баллов кандидатов

Отдельно признаки момента времени

- День недели
- Время суток

# Результаты

- Было до нас: (эвристический алгоритм)

$\text{recall}@1 = 63,7\%$ ;

$\text{recall}@2 = 78,5\%$ ;

$\text{recall}@3 = 84,6\%$

- Стало:

$\text{recall}@1 = 72,1\%$ ;

$\text{recall}@2 = 82,6\%$ ;

$\text{recall}@3 = 88\%$

# Резюме

- I. Что прогнозировать и как использовать модель
- II. Как оценивать качество (онлайн и оффлайн)
- III. Выборка, признаки и модель: насколько это важно
- IV. Примеры задач, где не всё просто

# Спасибо за внимание



[info@applieddatascience.ru](mailto:info@applieddatascience.ru)



[https://t.me/joinchat/B10lThC96v0BQCvs\\_joNew](https://t.me/joinchat/B10lThC96v0BQCvs_joNew)



[https://github.com/vkantor/ml2018jan\\_feb](https://github.com/vkantor/ml2018jan_feb)