Data Science для оптимизации бизнеса

От интернет-магазинов до заводов

О лекторе

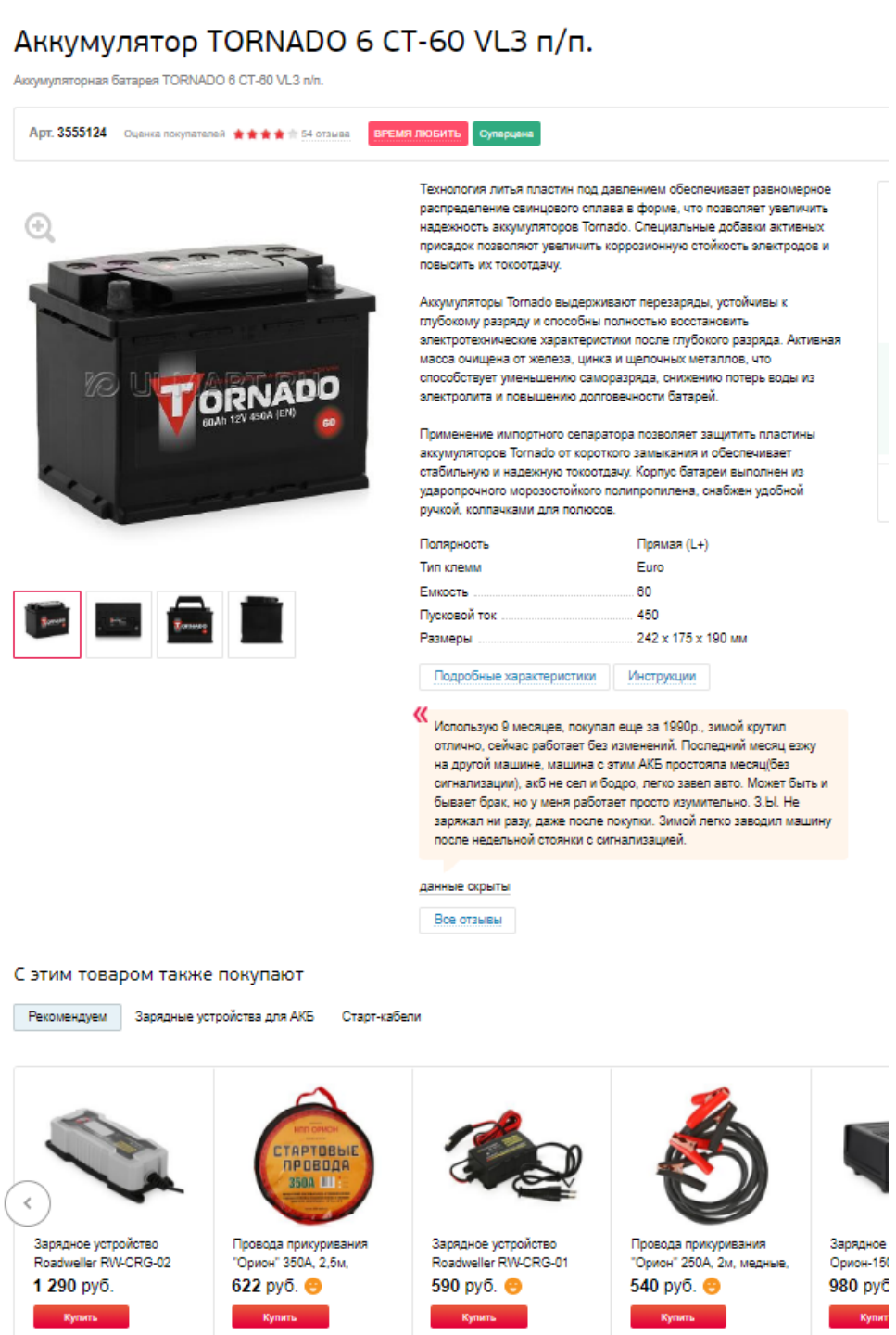
- Александр Белугин

- Преподаватель ФКН
- Yandex Data Factory
  - Руководитель проектного офиса
  - 50+ проектов

E-commerce

- Рекомендации товаров
- Например, на карточке товара
- «Похожие»
- «Часто покупают вместе»

- Как это работает без ML?



Дата показа	Код клиента	Код страницы	Код товара	Клик
				0/1

Дата показа	Код клиента	Код страницы	Код товара	Клик
				0,99
				0,98
				0,82
				0,81
				0,74
				0,72

Генерация признаков

- Дата
  - День недели
  - Праздник?
- Код товара
  - Категории
  - Параметры
  - Признаки из описания
  - Признаки из картинки
- Код клиента…

А/Б тест

- Дает ли больше денег система рекомендаций?

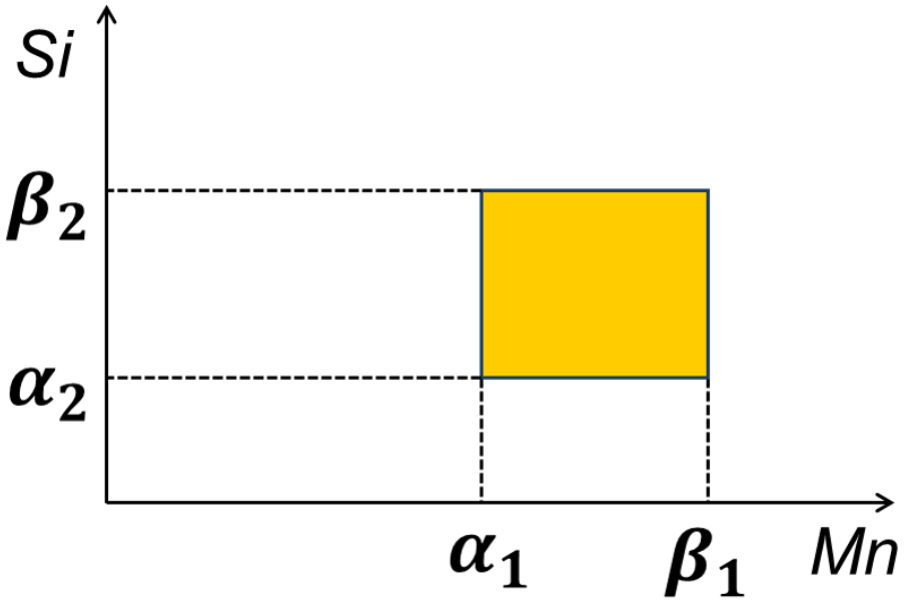
- Нужно правильно разбить на группы
  - Hash от username+соль \ число групп
  - Нужна контрольная группа
- Как проверить разбиение? АА-тест
- Ухудшающие эксперименты
- Перемешивание групп если тестов много

Проблемы на примере e-commerce

- Пользователь залогинился – что стало с его группой?
- Телевизор за 6,5 млн. рублей
- Тестовая закупка сотрудником
- B2B-заказ
- Параллельные тесты
- Ошибки Метрики и GA

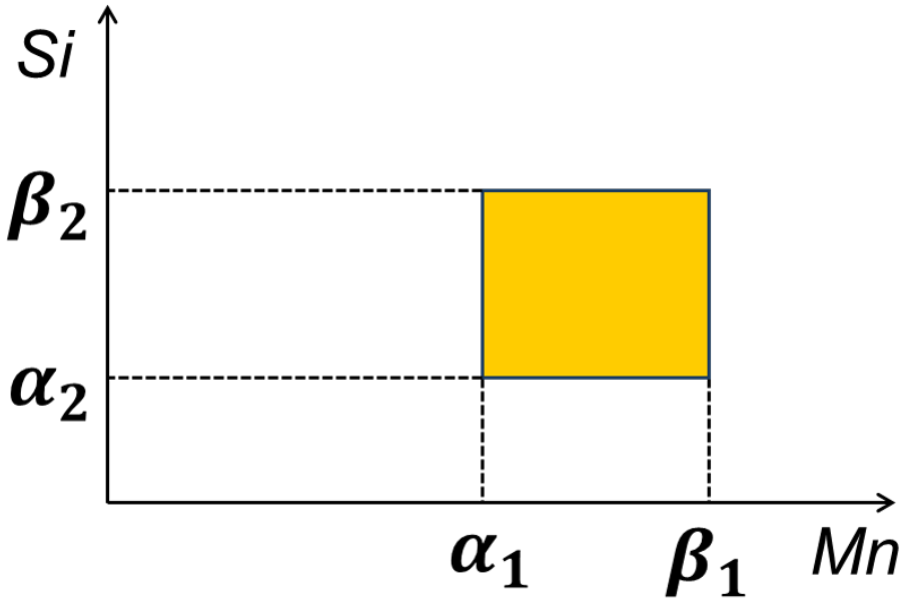
Много примеров

Оценка эффекта при производстве стали



- $$$$$$

Потенциал оптимизации



- $$$

Достигнутая экономия

ZARA
ZARA

Можно ли создать сериал на основе данных?

AN AMAZON ORIGINAL SERIES
ALPHA
HOUSE
5 NEWS
CHANNEL 3

Еще примеры

- Сеть знакомств и их главная страница
- Спрос на наличные в банкоматах: Москва и регионы
- Удержание клиента
- Предсказание снижения производительности электролизёров
- Сушка табака

Пример расчета — автомодерация

				Кол-во фотографий			Вариант "три класса"			
				Фото в час	2 700		Итого стоимость руб.		4 750 000	
				В смене чел.	5					
				Кол-во смен	4		Вариант "много классов"			
		-		Часов в смене	12		Итого стоимость руб.		6 250 000	
Цены, чтобы отбить затраты за 1 год				Фото в сутки:	648 000					
Тариф	Ставка, руб за 1000 фото	Итого в год		Фото в год:	236 520 000					
"Три класса"	20,08	4 750 000		Тысяч фото в год:	236 520					
"Много классов"	26,42	6 250 000								
				Затраты на модераторов			Доля фото где модель дает 95% точность			
Окупаемость заказчика	качество	макс.цена		Возможный оклад	28 000					
Максимальная цена	60%	25,57		Налоги, %	50%		Доля	Эффект	Прибыль заказчика	
	70%	29,83		Итого в месяц	840 000		60%	6 048 000	- 202 000	
	80%	34,09		Итого в год	10 080 000		70%	7 056 000	806 000	
	90%	38,36					80%	8 064 000	1 814 000	
							90%	9 072 000	2 822 000	

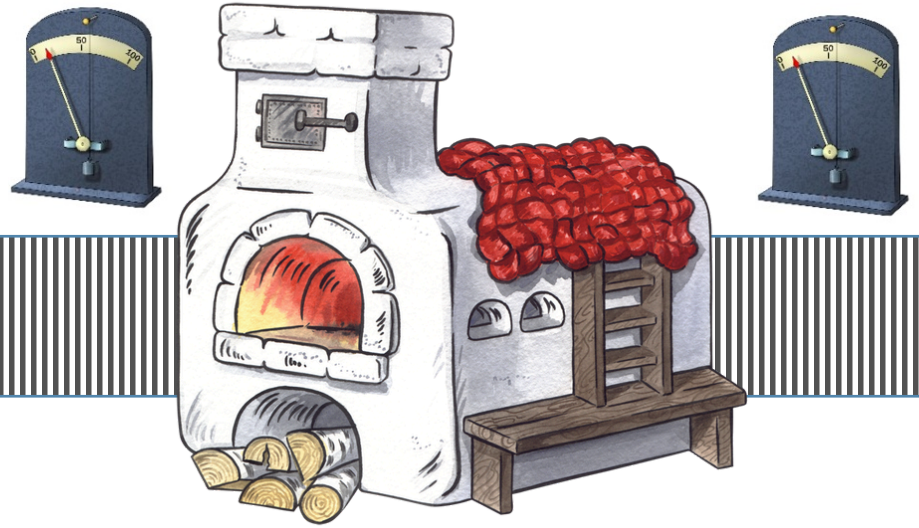
Еще примеры

- Сеть знакомств и их главная страница
- Спрос на наличные в банкоматах: Москва и регионы
- Удержание клиента
- Предсказание снижения производительности электролизёров
- Сушка табака

Слишком сухой табак

Табак ок

Слишком влажный табак



← 10 минут →

Образовательные проекты ВШЭ

Списком

- Data Culture
  - Откуда взять данные?
  - Что с ними можно делать?
  - Какие инструменты использовать?

- ДПО «Data Science для менеджеров»
  - Введение в машинное обучение и анализ данных
  - Постановка задачи и особенности управления проектами
  - Оценка качества проектов в Data Science
  - Построение моделей
  - Подготовка технического задания и управление проектом

Анализ изображений вместо переписи

Классификация машин из Google Street View (97% совпадения)

Обучение модели на малой части городов

Тест на других городах

Корреляции:

- 61% уровень краж
- 82% уровень дохода
- 66% уровень загрязнений

http://arxiv.org/abs/1709.02480

Коммерческие проекты ВШЭ

Исследовательская группа Байесовских методов

- 3 старших исследователя
- 6 аспирантов
- 11 студентов

- Совместно с ВМиК МГУ

Примеры проектов

- Предсказание дебита нефти для Шлюмберже
  - Результат зависит от выбора места скважин, способов воздействия, схем отвода воды и т.д.
  - Классические модели работают слишком долго и не учитывают всех данных
- Сбербанк и анализ клиентов
  - Машинное (векторное) представление каждого пользователя
  - Визуализация кластеров пользователей
  - Визуализация изменения предпочтений пользователей во времени
  - Использование в моделях скоринга и рекомендаций