

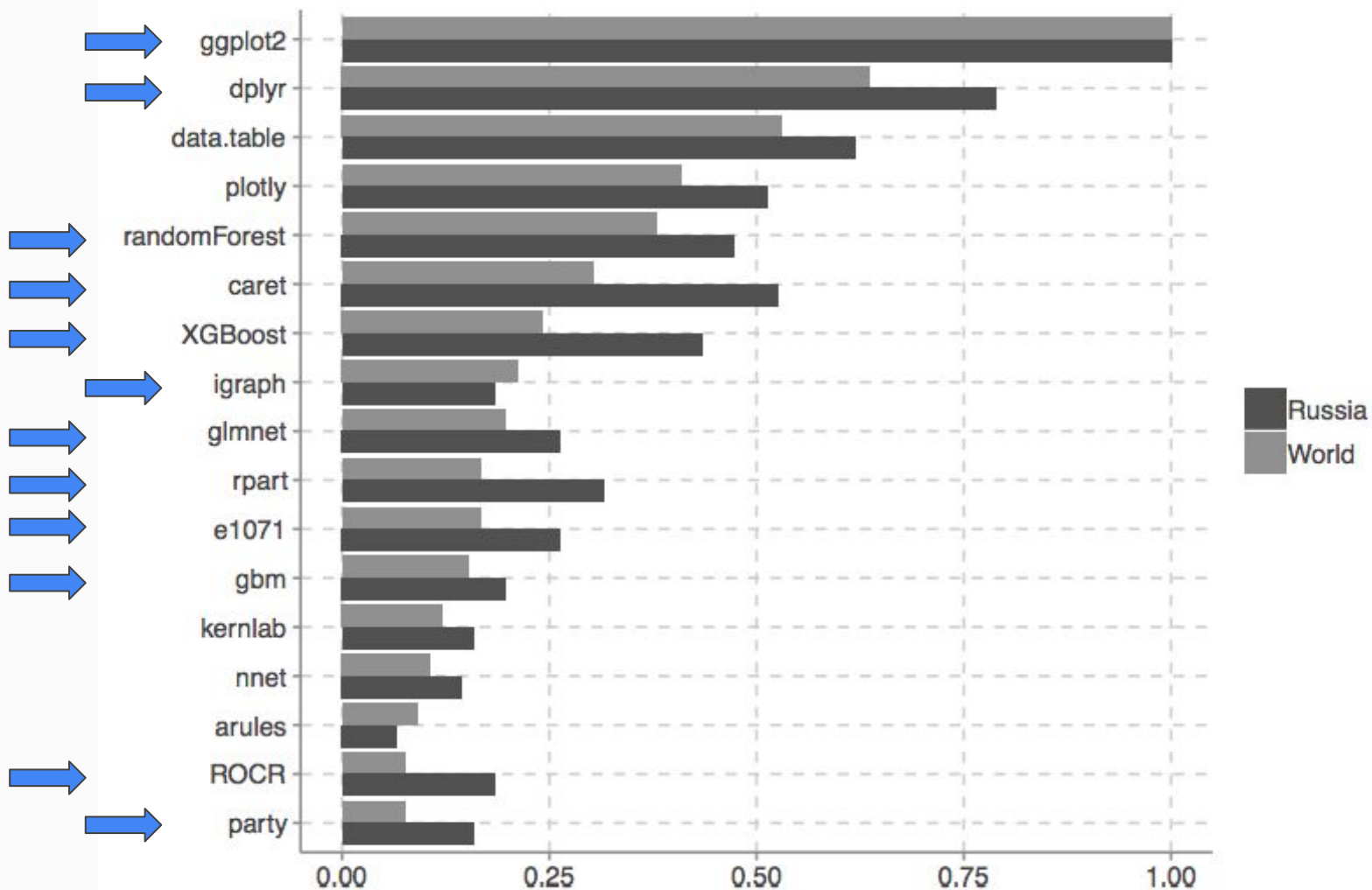


Приложения и практика анализа данных

Data Science Minor, зима-весна 2020

Что мы знаем и умеем

- Этапы анализа данных
- Основы агрегации данных
- Основные способы визуализации
- Немного про статистические тесты
- Рекомендательные системы
- Анализ текста
- Анализ сетей
- Машинное обучение

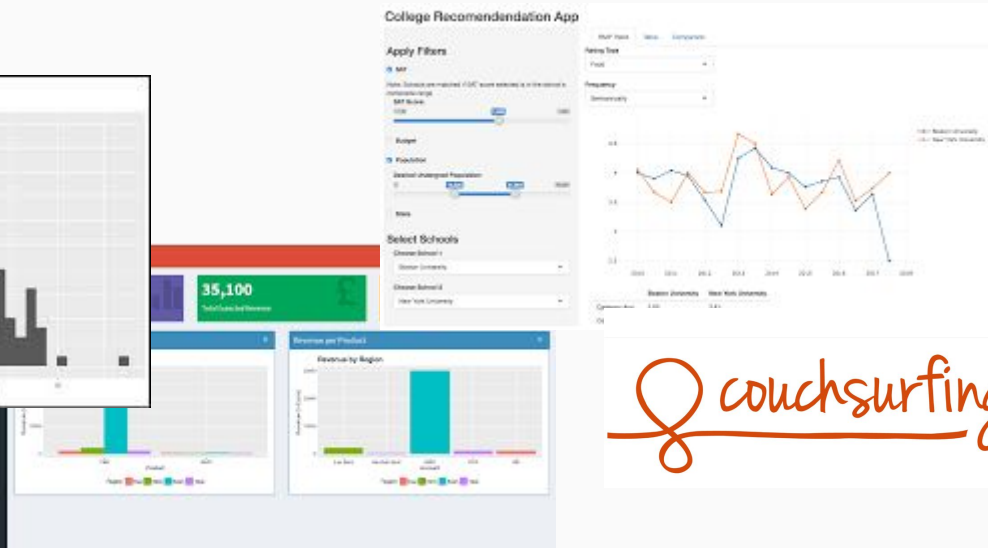


Еще про инструментарий

- Создание отчетов и RMarkdown (knitr)
- Работа с датами (lubridate)
- Немного работы со строками (stringr)
- Работа с текстами (tidytext, quanteda, topicmodels)
- Еще работа с сетями (sna + igraph)
- Рекомендации (recommenderlab)
- ...

Что мы будем делать в этом семестре

- Итоговый проект



Итоговый проект. Вариант 1.

- Рекомендательная система по любой тематике с интерфейсом
 - “с интерфейсом” = кнопки, списки, формы, т.е. пользователь может что-то изменить, приложение / сервис реагирует на действия пользователя
 - “рекомендательная” = в широком смысле рекомендации, т.е. советует, предлагает варианты, что-то оценивает

Итоговый проект. Вариант 2.

- Исследовательский проект с интерактивным отчетом
 - “с интерфейсом” = осмысленная визуализация результатов, уточнения разные формы, приложение / сервис реагирует на действия пользователя
 - “исследовательский” = проект построенный на осмысленном наборе данных с постановкой вопроса и ответом на него методами анализа данных

Итоговый проект: формально

Обязательно

- интерфейс (не отчет)
- взаимодействие с пользователем
- применение методов машинного обучения

Итоговый проект: совсем формально

$$\text{Орез} = 0.3 * \text{Одн} + 0.1 * \text{Оидея} + 0.3 * \text{Озащита} + 0.3 * \text{Оотчет},$$

где:

Одн – оценка за дневники проекта,

Оидея – оценка за защиту идеи проекта,

Озащита – оценка за итоговую защиту проекта,

Оотчет – оценка за итоговый проект.

Итоговый проект: инструментарий

- использование [Git](#) (контроль версий и совместная работа над проектом) настоятельно рекомендуется
- разработка -- на чем угодно
- базовая версия: R + Shiny

Итоговый проект: организационно

- Вариант 1 (инициативный):
 - придумать и согласовать идею
 - собрать команду (4-6 человек, не более 2 человек с одной ОП)
 - найти / собрать данные
 - выполнить проект
- Вариант 2 (стандартный):
 - данные предоставляем мы
 - на команды разбиваем случайно
 - нужно: оформить идею и выполнить проект

Итоговый проект: организационно 2

- Начало семестра -- новые темы
 - Git
 - Shiny
 - немного про сбор данных (принципы API и основы (совсем основы) html)
- много самостоятельной работы
- конец 3 модуля и 4 модуль -- занятия в форме консультаций и дополнительных тем

Много самостоятельной работы ([расписание](#))

Аудитория	Дата	1 пара	2 пара	3 пара	4 пара	5 пара	6 пара
ауд.№435	11 янв.					DS	DS
к.к.№3.1	18 янв.	DS гр1	DS гр1	DS гр2	DS гр2	DS гр3	DS гр3
к.к.№3.2		DS гр4	DS гр4	DS гр5	DS гр5	DS гр6	DS гр6
к.к.№3.1	25 янв.	DS гр5	DS гр5	DS гр6	DS гр6	DS гр4	DS гр4
к.к.№3.2		DS гр2	DS гр2	DS гр3	DS гр3	DS гр1	DS гр1
	1 февр.	Нет занятий					
	8 февр.	Нет занятий					
к.к.№3.1	15 февр.	DS гр3	DS гр3	DS гр1	DS гр1	DS гр2	DS гр2
к.к.№3.2		DS гр6	DS гр6	DS гр4	DS гр4	DS гр5	DS гр5
	22 февр.	Нет занятий					
	29 февр.	Нет занятий					
к.к.№3.1	7 мар.				DS гр4	DS гр5	DS гр6
к.к.№3.2					DS гр1	DS гр2	DS гр3
	14 мар.	Нет занятий					
к.к.№3.1	21 мар.				DS гр4	DS гр5	DS гр6
к.к.№3.2					DS гр1	DS гр2	DS гр3

Итоговый проект: оценивание

- Групповая часть:
 - промежуточная презентация
 - финальная презентация
 - работающее приложение
- Индивидуальная часть:
 - промежуточные отчеты по этапам
 - итоговый письменный индивидуальный отчет

Этапы проекта

1. Придумать идею.
2. Собрать команду
3. Создать репозиторий для совместной работы над проектом (GitHub)
4. Составить план действий
5. Распределить роли
6. Распределить задачи
7. Выполнить проект =))

+ Промежуточный и итоговый отчеты + Обсуждения

Обязательные пункты плана: 1. Данные

- Есть данные / нет данных
- Ссылка на них, если есть. Описание того, как и когда они будут получены, если их нет
- Описание данных: формат, объем, как их можно загрузить и как начать с ними работать; переменные, их описание, где его можно посмотреть
- Очистка данных: нужна / не нужна, кто ее делает, как все члены группы могут понять, какая версия является актуальной, и где взять актуальную версию

2. Планируемое представление результатов

- что рекомендуется или какой вопрос исследуется
- что представляет собой результат рекомендации или какие вы ожидаете результаты исследования
- что должен сделать / ввести пользователь, чтобы получить рекомендацию или какие аспекты исследования планируется визуализировать и как

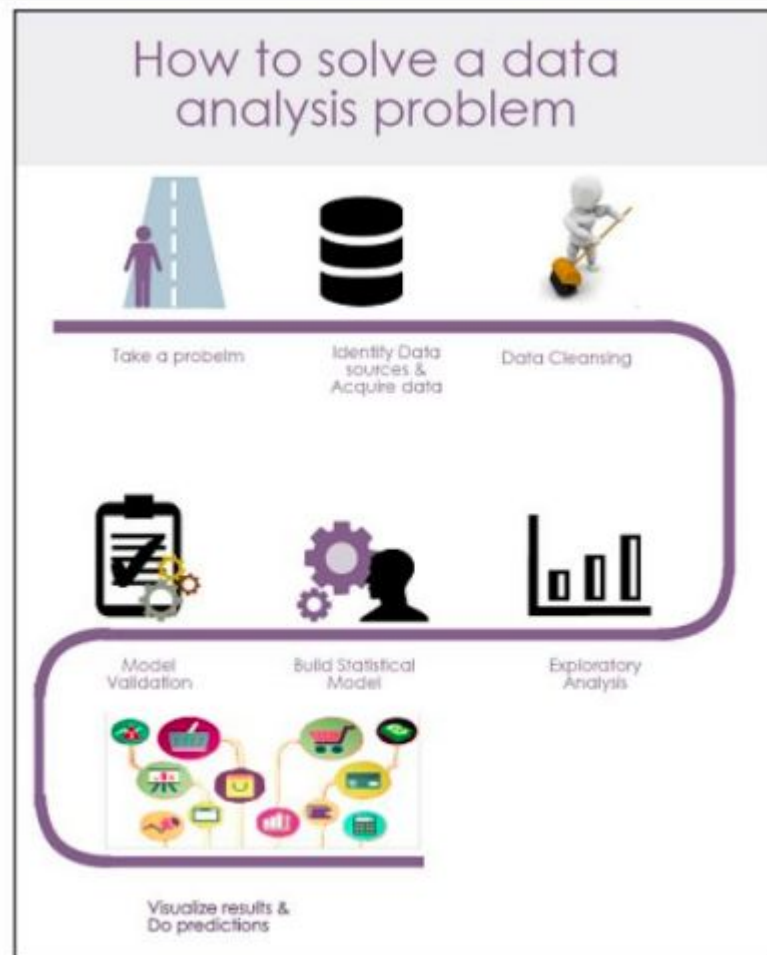
3. Методы

- какие методы и алгоритмы планируется использовать
- разведочный анализ: посмотреть основные закономерности

Обязательно: как минимум, один из методов машинного обучения, изученный в прошлом семестре (кластеризация, классификация, РСА...)

Роли

- По задачам
 - Сбор данных.
 - Обработка данных.
 - Построение моделей.
 - Визуализация результатов.
 - Общее руководство и пинание всех ногами.
- Все делают все...



Этапы (2)



Еще раз про рекомендательные системы

алгоритмы, которые пытаются предсказать, какие объекты (фильмы, музыка, книги, новости, веб-сайты) будут интересны пользователю, имея определенную информацию о его профиле.


























Пример:

- рекомендации в Amazon
- www.amazon.com
- зачем рекомендации? А как же "физические", "реальные" магазины?

Коллаборативная фильтрация

Используются известные предпочтения (оценки) группы пользователей для прогнозирования неизвестных предпочтений другого пользователя

Коллаборативная фильтрация

					
A					
B					
C					
D					
E					

Коллаборативная фильтрация: типы

- User-based

- Ищем пользователей наиболее похожих на нашего пользователя
- Рекомендуем то, что нравится им

- Item-based

- Выбираем наиболее понравившиеся пользователю предметы
- Ищем те предметы, что больше всего похожи по оценкам других пользователей на эти

Не только коллаборативная фильтрация

- схожесть
- свойства, характеристики
- кластеризация

Куда применить то, что мы изучили

Куда применить то, что мы изучили

- В вашей основной области:
 - агрегация,
 - визуализация
 - отчеты
 - автоматизация каких-то действий
 - сети и взаимосвязи
- Специализация в Data Science

Специализация в Data Science

[Data Analyst](#)

[Data Science](#)

Специализация в Data Science: что еще

- Базы данных и SQL (см. datacamp.com и codecademy.com)
- Python
- Большие данные (MapReduce, Spark...)
- Kaggle