

Курс МАДМО по data science

The background is a solid teal color. It features several faint, semi-transparent decorative elements: a large donut chart in the upper right, several smaller pie charts scattered around, and a bar chart in the bottom right corner.

Вводное занятие



План

- Знакомство
- План курса
- Какие задачи решает data science
- Почему python
- Финальный проект



О себе

- Марков Александр
- Сотрудник исследовательской группы байесовских методов в машинном обучении; лаборатории ВШЭ и Samsung; сотрудник МФТИ
- Карьера:
 - 3 года работы в области deep learning (обработка текстов, генеративные модели, ансамблирование, мета-обучение)
 - Yandex, Samsung AI Center
 - Опыт преподавания 2 года (МФТИ, ШАД)
- Выпускник ФИВТ МФТИ





Контакты

- Александр
 - telegram: @amarkv
 - почта: markov.av@phystech.edu
 - канал для объявлений: https://t.me/joinchat/SsGHWeWJmQZ_E6Tk
 - чат курса: <https://t.me/joinchat/Tf04lWTE5UqshniZ>
- Кураторы курса
 - Анастасия tg: @AnastasiaIvanovaMarketing
 - Дарья Клименко DAnKlimenko@sberbank.ru, tg: @DAKlimenko



Полезные ссылки

1. telegram: @amarkv
2. почта: markov.av@phystech.edu
3. git: <https://github.com/markovalexander/MADMO>
4. zoom: <https://zoom.us/j/99158794795?pwd=M2tSU2RMUHN3bUg1K21LNGpuR041QT09>
5. чат: <https://t.me/joinchat/Tf04IWTE5UqshnjZ>
6. новости: https://t.me/joinchat/SsGHWeWJmQZ_E6Tk
7. анкета-знакомство: <https://forms.gle/46JHLnWbEegUMn2k8>



План занятий

Занятия проходят по вторникам и четвергам, время занятий: 18:30 - 21.30
Всего 8 занятий. На занятиях 9-10 будет сдача проекта

Примерное расписание:

18:30 - 19:25 - занятие

перерыв 5 мин

19:30-20:40 - занятие

перерыв 5 мин

20:45-21.30 - занятие

итого - 3 часа в день



Темы (примерные)

- 1) *Python3
 - 2) Работа с данными в python3
 - 3) Математика и визуализация данных в python
 - 4) Линейная регрессия
 - 5) Очистка и подготовка данных
 - 6) Метрики регрессии
 - 7) Классификация, логистическая регрессия
 - 8) Метрики классификации
 - 9) Feature Engineering
 - 10) SVM
 - 11) Деревья решений
 - 12) Подбор гиперпараметров
- Тестирование**
Сдача проектов, консультация



Оценка

оценка за курс выставляется по следующей формуле:

$$O = [(Дз1 + Дз2 + Дз3)/3 + ТЕСТ + ПРОЕКТ]/3$$

- на каждую домашнюю работу отводится ровно одна неделя
- жестким дедлайном считается начало семинара через неделю
- сдача проекта обязательна для получения сертификата
- для получения зачета необходимо набрать 45% баллов
- для получения зачета с отличием 🧐 необходимо набрать 80% баллов



Организационные моменты

- Домашние задания: задачи для повторения и закрепления материала, пройденного на занятиях, в формате юпитер ноутбука
- Вопросы на занятиях задавать можно и нужно
- Обратная связь в конце каждого занятия (гугл-форма на 2 минуты)
- Общение происходит в Telegram
- Задания сдаются на почту с темой вида “[МАДМО] Фамилия Задание 1”

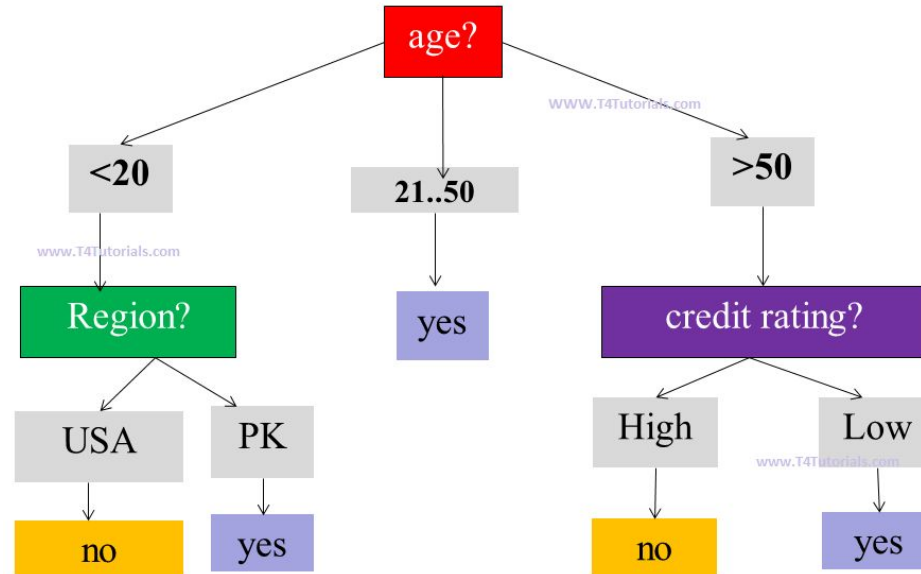
Для чего нужен DS

Пример: экспертная система. Выдать кредит человеку или нет?

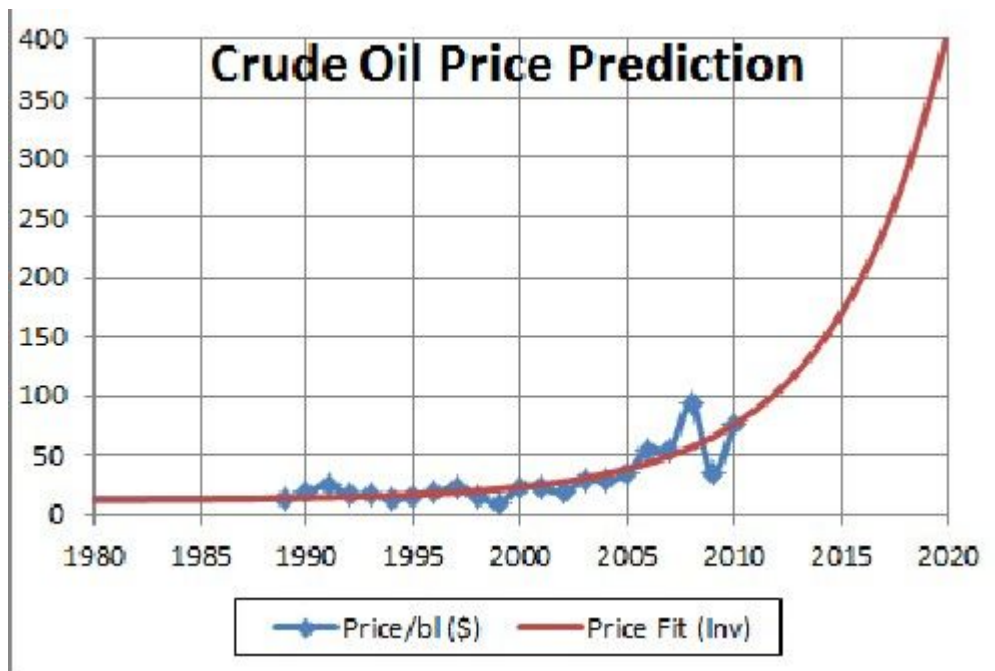
Так выглядят исторические данные:

Age	Job	Marital Status	Education	Has credit in default	Avg. credit balance	Has housing loan	Has personal loan	Contact type	Last contact day	Last contact month	Last contact duration (sec)	Number of contacts	Days passed	Previous contacts	Outcome previous campaign	Subscribed deposit
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no

Так может выглядеть автоматически построенная модель:



Еще примеры



Специально для Вас

Аффинаж, rsac, номер скрыт и другие

▶ СЛУШАТЬ ВСЕ



Google

Приглядитесь к этим предложениям



3 295 ₽ **-50 %**
6 590 ₽

Кеды VANS



7 030 ₽ **-35 %**
10 800 ₽

Внешняя звуковая



1 875 000 ₽
Виниловый
проигрыватель Spira...



4 400 ₽ **-30 %**
6 290 ₽

Кеды VANS



11 790 ₽
Лонгборд GoldCoast
Standard

где найти|

где найти
где найти **работу**
где найти **девушку**
где найти **друзей**
где найти **парня**
где найти **мужа**
где найти **деньги**
где найти **ответы на огэ 2018**
где найти **ответы на егэ 2018**
где найти **алису**

кто лучше сири или ты

Я сейчас чуть не удалилась в сердцах. Конечно, я не Сири.

переведи меня зовут алиса на английский

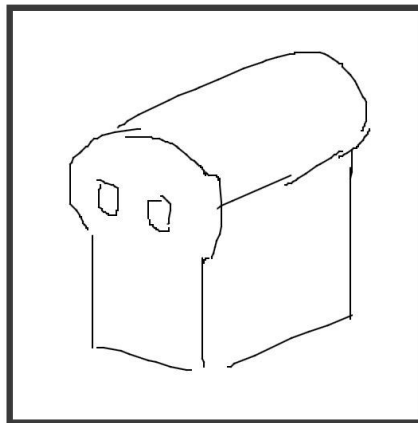
Лет ми спик фром май харт:
пойдёмте в Яндекс.Переводчик.
Там переводить проще.

[ОТКРЫТЬ ЯНДЕКС.ПЕРЕВОДЧИК](#)

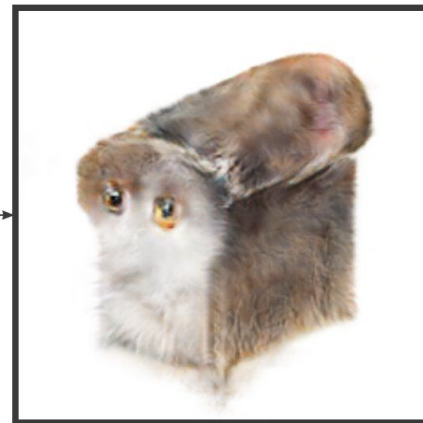
[ПОИСКАТЬ В ЯНДЕКСЕ](#)



INPUT



OUTPUT



pix2pix

process

Язык python



В data science и в этом курсе используется язык программирования Python.

Первое занятие будет посвящено изучению инструментов python для работы с данными

Рекомендуется самостоятельно попрактиковаться в программировании на python. Хороший интерактивный курс: <http://pythontutor.ru/>



Почему python

- Простота
- Из коробки
- Богатая библиотека
- Распространенность

На Python пишут в Яндексе, Гугле, Авито, Мэйле, Касперском и даже в NASA.

Питон используется в Dropbox, Youtube, Bitly, Reddit, а Instagram, Disqus, Spotify, Bitbucket работают на Django! И таких примеров много. А вот ещё один факт:

Python занял третье по популярности место на GitHub в 2016-м году



Программы на компьютер

Для занятий понадобятся только:

- Браузер
- Телеграм
- Гугл-аккаунт (почта)
- Регистрация на <https://www.kaggle.com/>
- Регистрация на <https://github.com/>

Для работы потом:

- [Anaconda](#) или `pip` или [miniconda](#)
- PyCharm или VS Code

Финальный проект

В конце курса слушателям предлагается сделать свой проект по data science. Это может быть:

- a) Рабочий проект. Применить методы, пройденные в курсе для своей задачи и на своих данных.
- b) Соревнование на <https://www.kaggle.com/>. Построить модель на реальных данных и посоревноваться с людьми со всего мира.

Необходимо зафиксировать тему финального проекта к концу третьей недели курса, иначе проект считается не сданным

Финальный проект

Примеры из прошлых курсов:

- Предсказать время отказа банкомата
- Кластеризовать банкоматы на территории России
- Предсказать нагрузку на центр поддержки в определенный день
- Купит ли клиент услугу, если ему показать уведомление?
- и многое другое...



Полезные ссылки

1. telegram: @amarkv
2. почта: markov.av@phystech.edu
3. git: <https://github.com/markovalexander/MADMO>
4. zoom: <https://zoom.us/j/99158794795?pwd=M2tSU2RMUHN3bUg1K21LNGpuR041QT09>
5. чат: <https://t.me/joinchat/Tf04IWTE5UqshnjZ>
6. новости: https://t.me/joinchat/SsGHWeWJmQZ_E6Tk
7. анкета-знакомство: <https://forms.gle/46JHLnWbEegUMn2k8>

