Школа data science

Вводная лекция

План

- Преподаватель
- План занятий
- Для чего нужен data science
- Почему python
- Финальный проект

О себе

- Руслан Байназаров
- Выпускник МГУ
- Сотрудник лаборатории инноватики МФТИ
- Карьера:
  - 2 года работы в области deep learning и обработки текстов, чат-боты
  - научная работа в NLP
- Контакты:
  - telegram: @nfthl
  - почта: hocop@yandex.ru

План занятий

1) Python3
2) Работа с данными в python3
3) Математика и визуализация
4) Регрессия
5) Очистка и подготовка данных
6) Метрики регрессии
7) Классификация, простые модели
8) Feature engineering
9) Метрики классификации
10) Логистическая регрессия
11) SVM классификация
12) Checkpoint. Kaggle
13) Деревья, ансамбли
14) Boosting, Stacking
15) Обработка текстов
16) Дальнейшее развитие

Тестирование

Сдача проектов, консультация

План занятий

4 учебных дня. Пятый - сдача проектов

Расписание:
9:30 - 11:00 - занятие
перерыв 20 мин
11:20-12:30 - занятие
обед 1 час
13:30-15.00 - занятие
перерыв 20 мин
15.20-16.30 - занятие
перерыв 20 мин
16:50-18:00 - занятие

итого - 7 часов в день

План занятий

- Домашние задания - Простые задачи для повторения и закрепления пройденного на занятиях
- Вопросы на занятиях - задавать можно и нужно
- Обратная связь в конце каждого занятия (2 минуты)

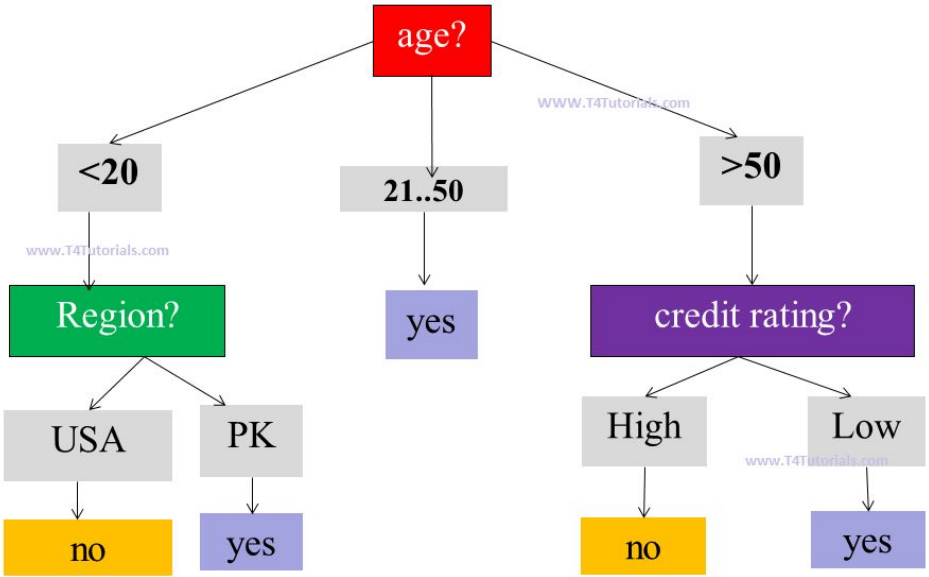
Для чего нужен DS

Пример: экспертная система. Выдать кредит человеку или нет?

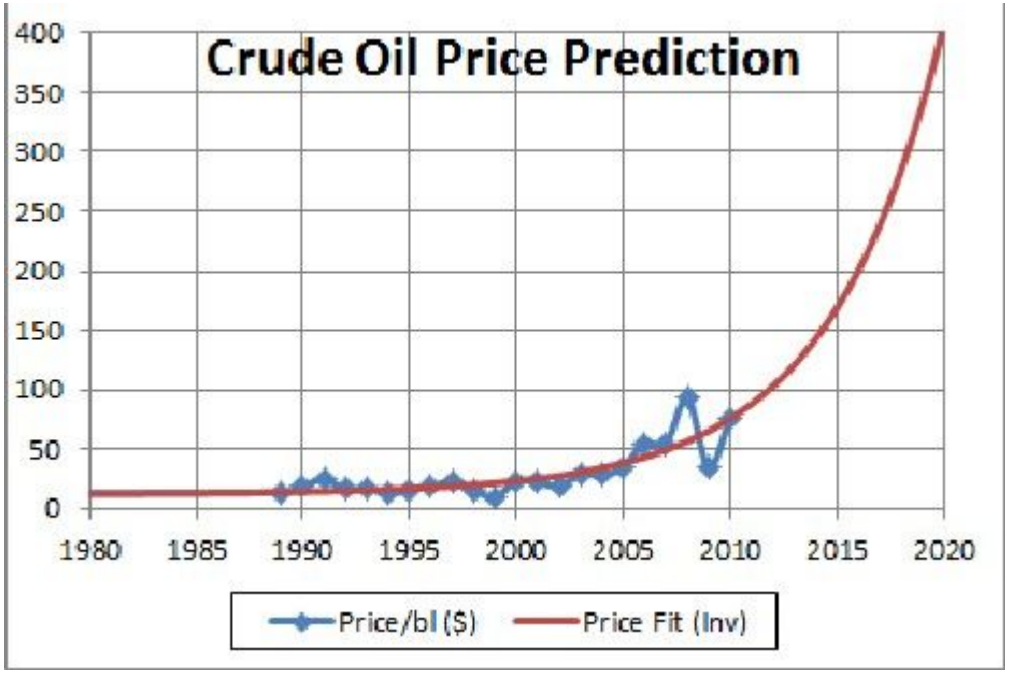
Так выглядят исторические данные:

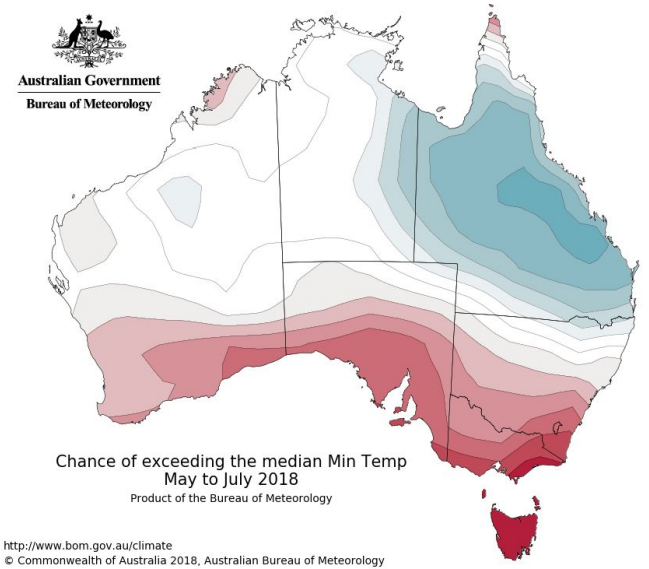
Age	Job	Marital Status	Education	Has credit in default	Avg. credit balance	Has housing loan	Has personal loan	Contact type	Last contact day	Last contact month	Last contact duration (sec)	Number of contacts	Days passed	Previous contacts	Outcome previous campaign	Subscribed deposit
30	unemployed	married	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


Crude Oil Price Prediction
400
350
300
250
200
150
100
50
0
1980
1985
1990
1995
2000
2005
2010
2015
2020
Price/bl ($)
Price Fit (Inv)


Australian Government
Bureau of Meteorology
Chance of exceeding the median Min Temp
May to July 2018
Product of the Bureau of Meteorology
80
75
70
65
60
55
50
45
40
35
30
25
20
Chance of exceeding median min temp (%)
http://www.bom.gov.au/climate
© Commonwealth of Australia 2018, Australian Bureau of Meteorology
Issued: 12/04/2018
Model Run: 08/04/2018
Base Period: 1981-2010

Сначала интересные

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

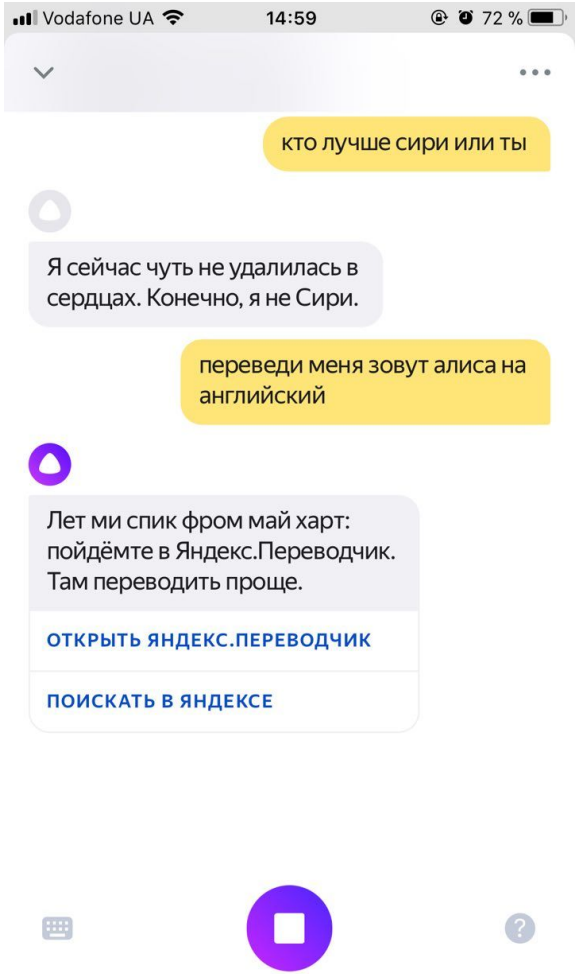
1 875 000 ₽
Виниловый проигрыватель Spira...

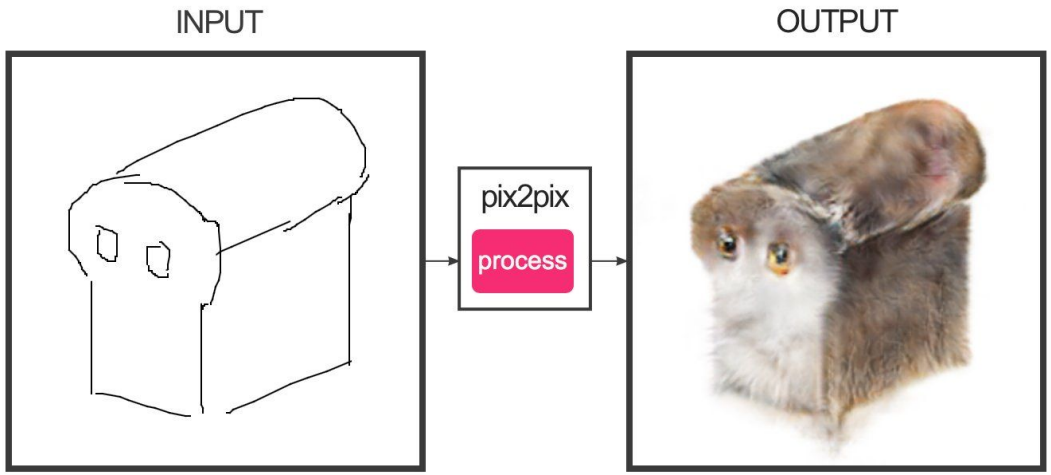
4 400 ₽ -30 %
~~6 290 ₽~~
Кеды VANS

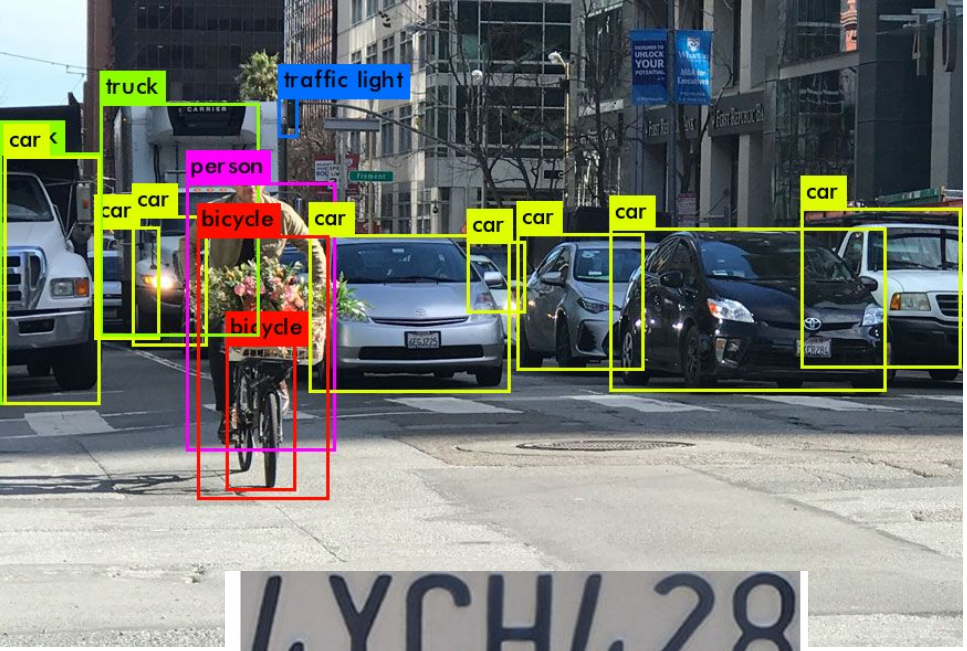
11 790 ₽
Лонгборд GoldCoast Standard

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**


Vodafone UA
14:59
72 %
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


truck
traffic light
car
person
car
car
bicycle
car
car
car
car
car
bicycle


Fetch! FUN
Brittany?
Sporty, strong desire to please, sensitive & devoted
© Getty Images/Microsoft Fetch


4YCH428
4YCH428
4YCH428


35
42
30
34
39

Язык python

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

Первые занятия будут посвящены python’у.

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, a 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

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.

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...