Кейс: (DVHB) Анализ схем севооборота

Команда №1 "Data Rescue Rangers"

Наша команда

Александр Максименко
г. Челябинск

Инженер-проектировщик РЗиА.
Активно занимаюсь изучением разработки нейросетей и ИИ , в дальнейшем хочу применять их в сфере энергетики.

Евгения Уварова
штат Нью-Джерси

Роль в команде
Project Manager

Имею опыт работы в международных компаниях и межконтинентальной релокации.

Андрей Ветров
г.Москва

Роль в команде
Product Manager

Образован в МИФИ как физик-теоретик в 1999, а нашел себя как исследователь данных в 2015. Увлекаюсь всем красивым: математикой, физикой, природой, женщинами.

Николай Филиппов
г. Москва

"Швец. Жнец.
На дуде игрец"
Всю жизнь в телекоме. Если у вас не ловит — это я виноват.

Алексей Долгополов
г. Москва

Работал в консалтинге.
"Работаю DA, но при этом хочу освоить еще и DS"

Дарья Трофимчк
г. Москва

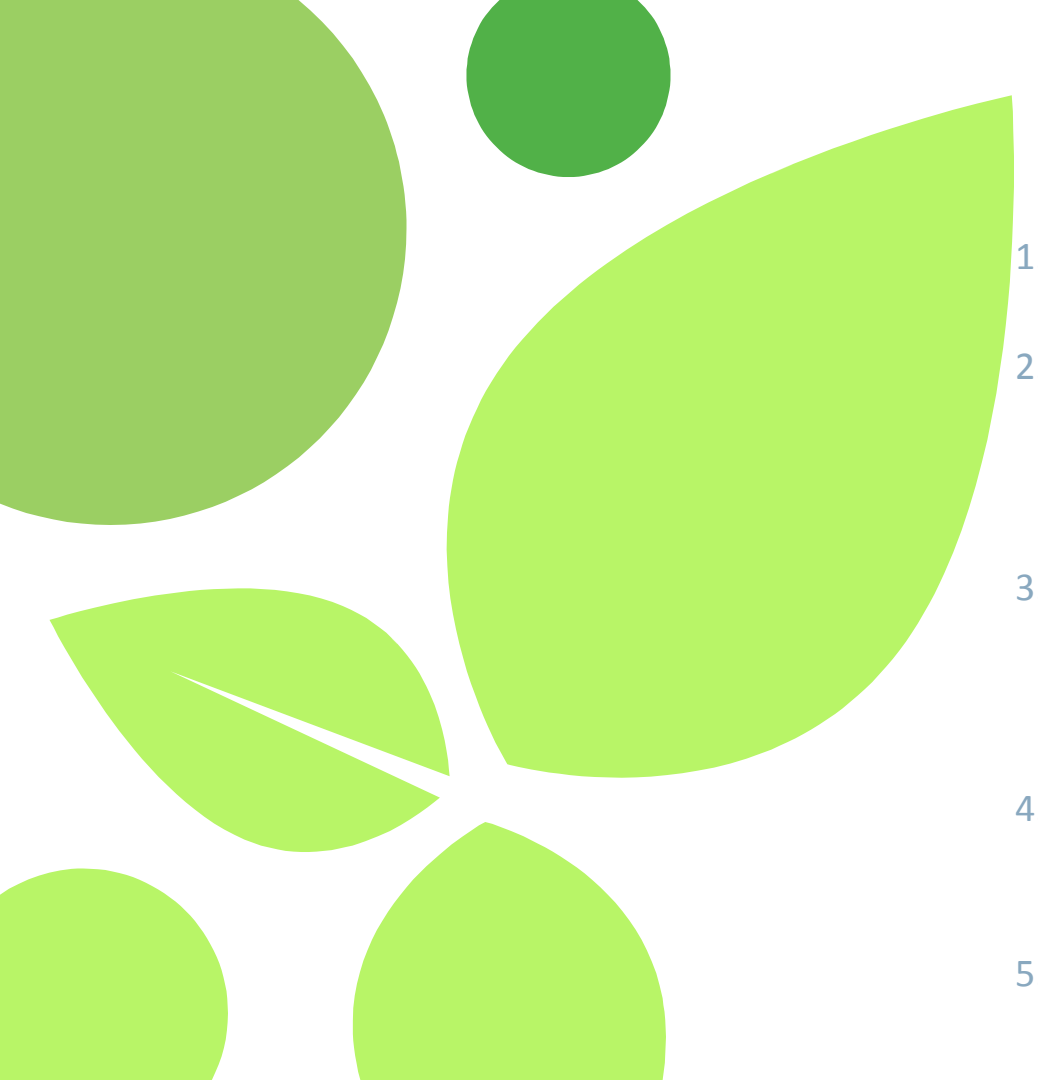
Не доверяю новостям, потому что знаю, как они пишутся. Учусь DS, чтобы сменить сферу деятельности.

Задача

На основе реальных исторических данных определить стратегии севооборота и предсказать культуры, выращиваемые на заданных полях в следующем году.

Необходимо:

- проанализировать набор данных
- выявить последовательность смены посевных культур на заданных полях за пять лет
- предсказать, какой культурой засеяли эти поля в 2020 году



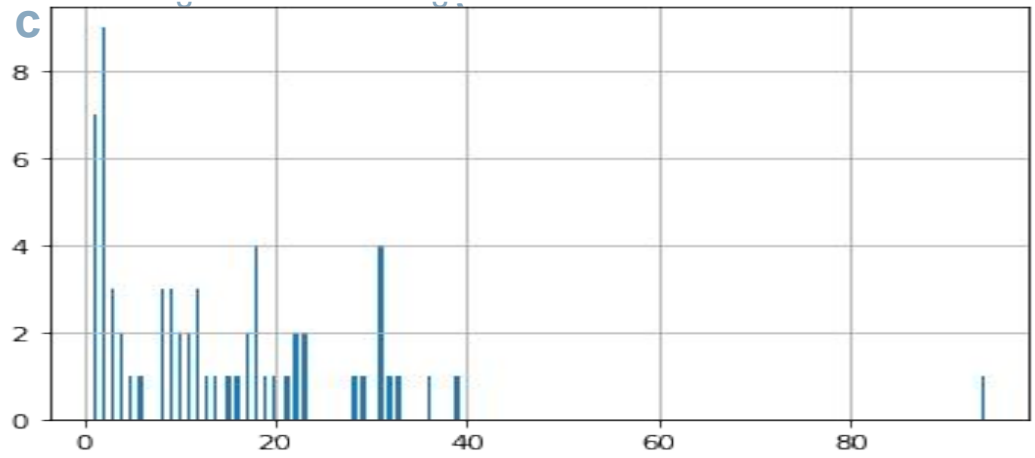
Гипотезы

1. Возможно выделить кластеры, содержащие ограниченное количество культур.
2. Для каждого из таких кластеров возможно обучить модель для предсказания культуры, которая будет засеяна на полях кластера в следующем году.
3. Список признаков можно обогатить дополнительными данными, в том числе сведениями об административных границах Франции, климатических зонах и тд.
4. На части полей регулярно сеется только одна культура, ее можно предсказывать по умолчанию
5. Возможно добиться точности предсказания (accuracy) > 80%.

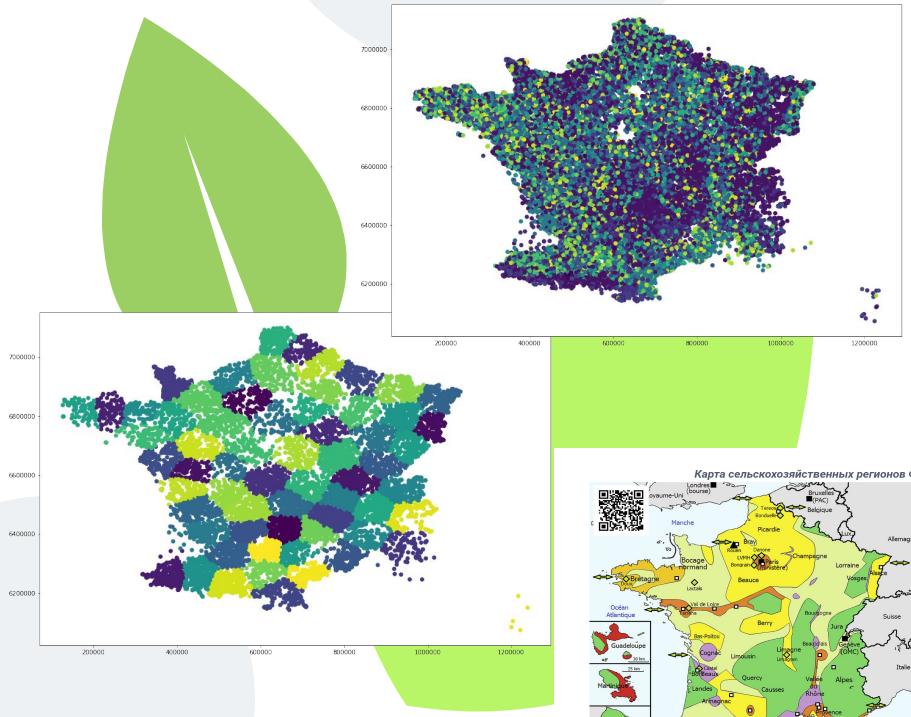
Принципы кластеризации

- Географическая
- По составу засеваемых культур
- По группам культур

Гипотеза 1 подтвердилась: найдены кластеры, содержащие ограниченное количество культур (в том числе кластеры с



География или посевы?



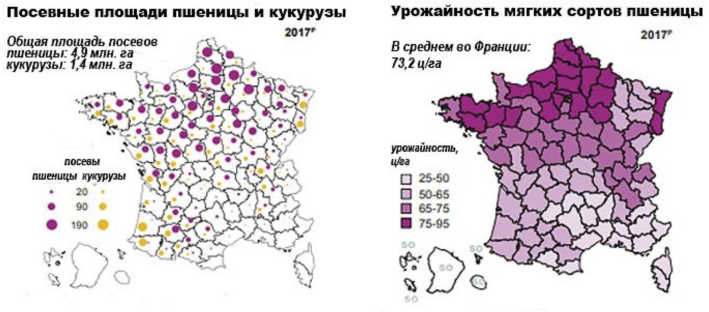
Кластеры на основе состава засеваемых культур плохо коррелируют с географическими, но

- адекватно отражают распределение сельскохозяйственных регионов Франции;
- есть 16 кластеров (почти половина наблюдений), содержащих одну культуру.

Разумный подход - кластеризация на основе кодов культур и использование географического кластера в качестве признака.

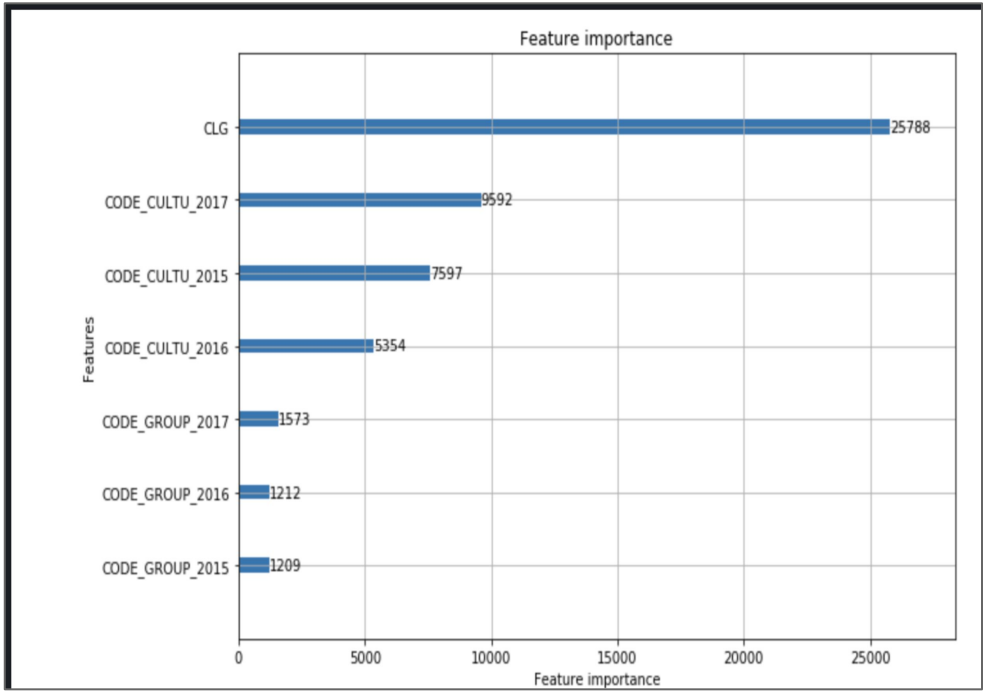
Карта сельскохозяйственных регионов Франции

Посевные площади и урожайность пшеницы и кукурузы во Франции (2017 г.)



Признаки для классификации

Наиболее значимым признаком оказался **географический кластер**.



Момент истины

В качестве классификатора выбран LightGBM - алгоритм градиентного бустинга, разработанный компанией Microsoft.

- Обучаемся на 2015-2017 из train. Target - 2018.
  Тестируемся на данных из test_2019.

```
In [ ]: accuracy_score(test_2019['CODE_CULTU
Out[294]: 0.8587882333901828
```

- Обучаемся на 2015-2018 из train. Target - 2019.
  Тестируемся на данных из test_2020.

```
In [ ]: accuracy_score(test_2020['CODE_CULT
Out[182]: 0.8148148148148148
```

Подтвердилась ли гипотеза 5? Скоро узнаем.

Проблемы

1. Мало времени. Мы, например, не успели проверить гипотезу 3.
2. Мало вычислительных мощностей. Некоторые идеи просто не получилось реализовать на домашних компьютерах и в Colab.
3. Мы умеем не все, что хотели бы сделать. Мы, конечно, научимся, но это займет какое-то время. GOTO п. 1.

Хотели, но...

1. Географическая кластеризация по границам административно-территориального деления (регионы, департаменты).
2. Географическая кластеризация по климатическим зонам.
3. Кластеризация по десятилетним последовательностям групп культур с учетом цикличности.
4. Ансамбли классификаторов.
5. И другие идеи разной степени безумности.

Ссылки

Открытые источники

https://www.data.gouv.fr/en/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-ilots-culturaux-et-leur-groupe-de-cultures-majoritaire/

Гитхаб

https://github.com/delhian/01_hak_9_03

Колаб

https://colab.research.google.com/drive/1UbUziZCwEE0qR_aZcc3L-1BZ2UGXCsma

Спасибо! Спрашивайте!