

# Прогнозирование урожайности подсолнечника по данным гидрометеорологической информации

Дипломный проект по профессии Data scientist

**Олег Воропаев**

Отраслевой аналитик, к.э.н.



# Цели и задачи дипломного проекта

1. Разработка web-скраперов для автоматизации процесса сбора входных данных
2. Создание базы данных для хранения собранной информации на платформе PostgreSQL
3. Разведочный анализ, предобработка, очистка и подготовка данных для обучения и тестирования модели
4. Разработка предиктивной модели для прогнозирования урожайности подсолнечника



# Постановка задачи. Современные методы оценки урожайности.

1



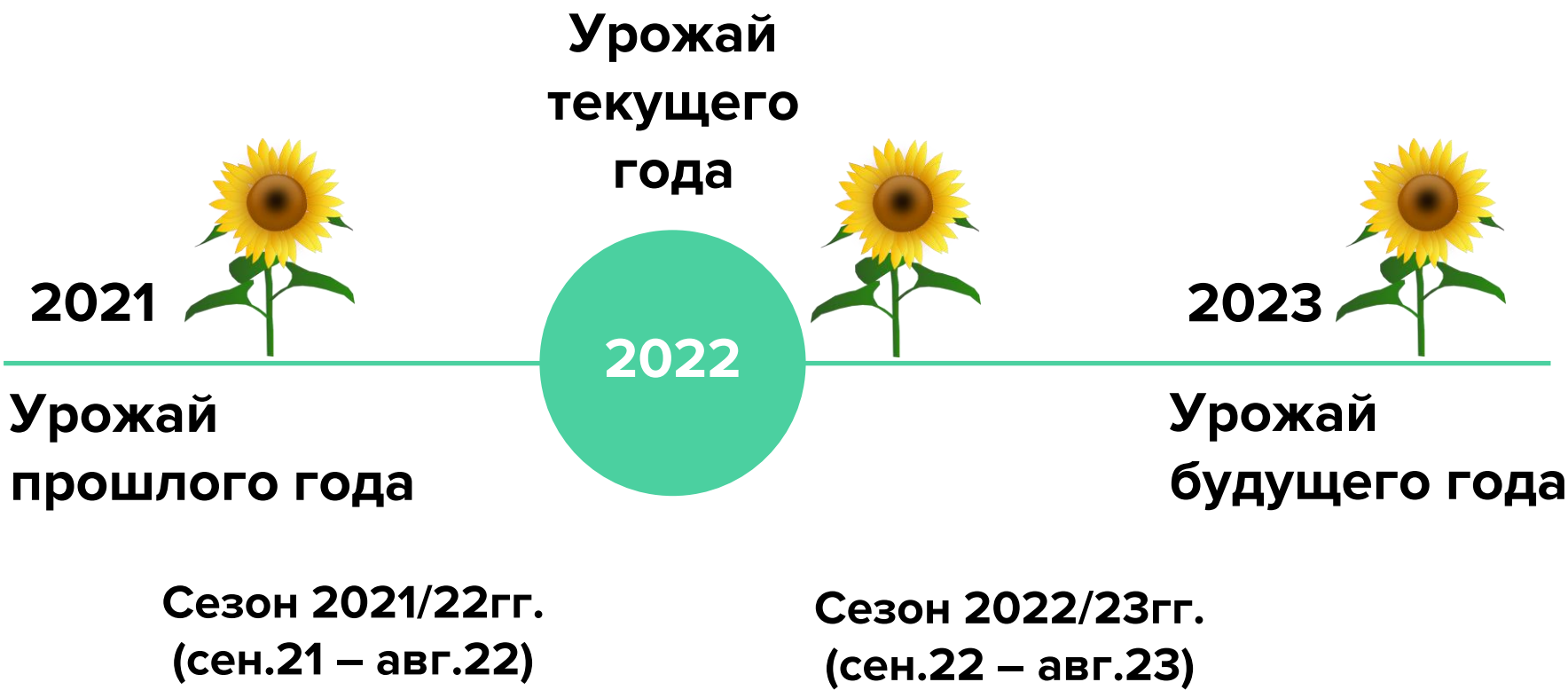
# Актуальность задачи

## Точный и своевременный прогноз определяет стратегию работы на весь сезон!

Особенность отрасли растениеводства – получение продукции (валового сбора) один раз в сезон.

**Главная задача аналитика (сгор-аналитика)** – своевременное и точное прогнозирование валового сбора перед началом сезона\*.

### Цикл производства подсолнечника



Объем валового сбора оказывает сильное **влияние на динамику цен и задает направление тренда** в течение всего сезона.

В случае сильного снижения валового сбора – возникает **дефицит продукции**, в случае сильного роста – появляется **избыток предложения**.

### Динамика средних цен и валового сбора подсолнечника в России

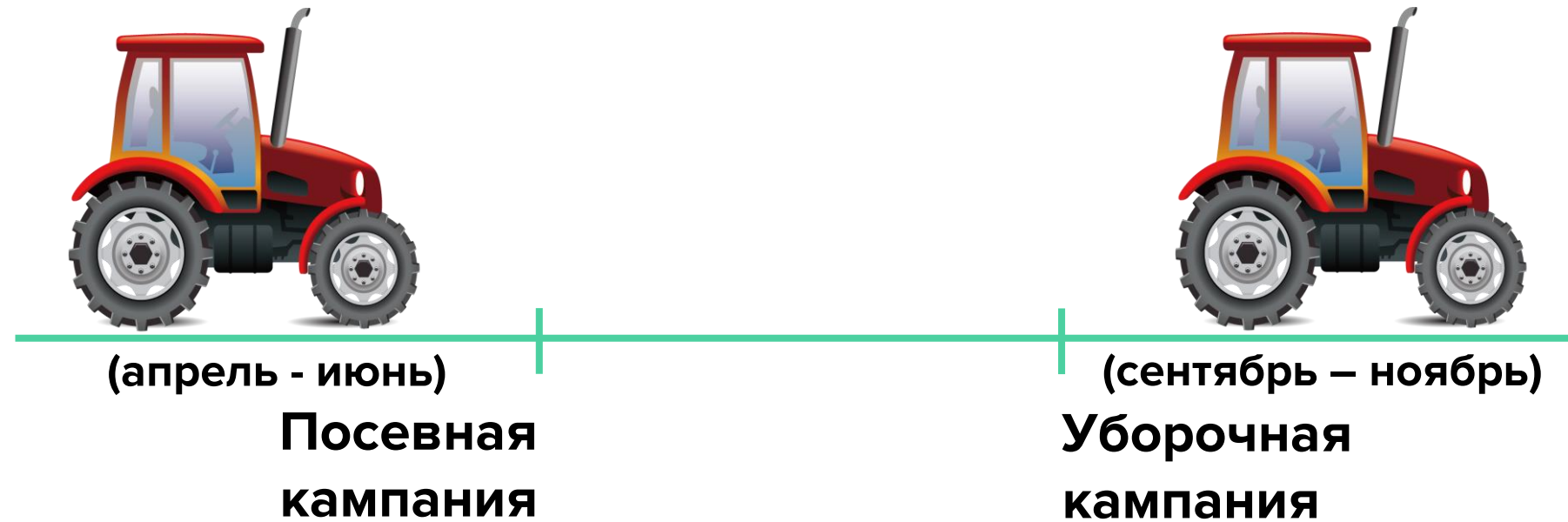


\* Начало и окончание сельскохозяйственного сезона не совпадает с календарным годом (для подсолнечника нач. сезона сентябрь тек. года – окончание август след. года).

# Актуальность задачи

Почему прогнозирование урожайности?

**Валовой сбор = Посевная площадь × Урожайность**



Старт посевной кампании начинается в апреле, завершается в конце июня\*.

**На 1 июля уже есть представление о размерах посевных площадей**

На начало сезона неизвестной переменной остается показатель урожайности

**К концу августа – началу сентября большая часть растений достигает фазы «восковой спелости», что позволяет проводить предварительную оценку урожайности**

\* Здесь и далее речь идет о выращивании подсолнечника





# Альтернативные методы оценки урожайности сельскохозяйственных культур

1

**КРОП-ТУР.** *Кроп-тур* - это выезд специалистов в поля с целью оценки качества и количества будущего урожая. Оценка, урожайности осуществляется лабораторным способом (замеры, взвешивания, визуальная и органолептическая оценка растений и урожая)



**Преимущества:** профессиональная оценка урожайности; можно увидеть то, что «не видно из кабинета»; широкий охват различных регионов (если маршрут кроп-тура продолжительный).

**Недостатки:** высокие финансовые затраты на организацию поездки (оплата работ, оборудования, транспорта); необходимо точно угадать со сроками поездки (пораньше-плохо, попозже еще хуже); субъективность специалистов.

Источник: личный фотоархив автора





# Альтернативные методы оценки урожайности сельскохозяйственных культур

2

**ЭКСПЕРТНАЯ ОЦЕНКА.** *Экспертная (кабинетная) оценка* – это методы, базирующиеся на опыте и знаниях отраслевых специалистов. Способы прогнозирования могут быть различными (экстраполяция, метод аналогов, мониторинг, «танцы с бубном» и т.д.). Качество и точность прогноза зависят от опыта эксперта.



**Преимущества:** низкие затраты; быстрота; авторские методики оценки могут быть очень эффективными; высококлассный эксперт = точный и качественный прогноз.

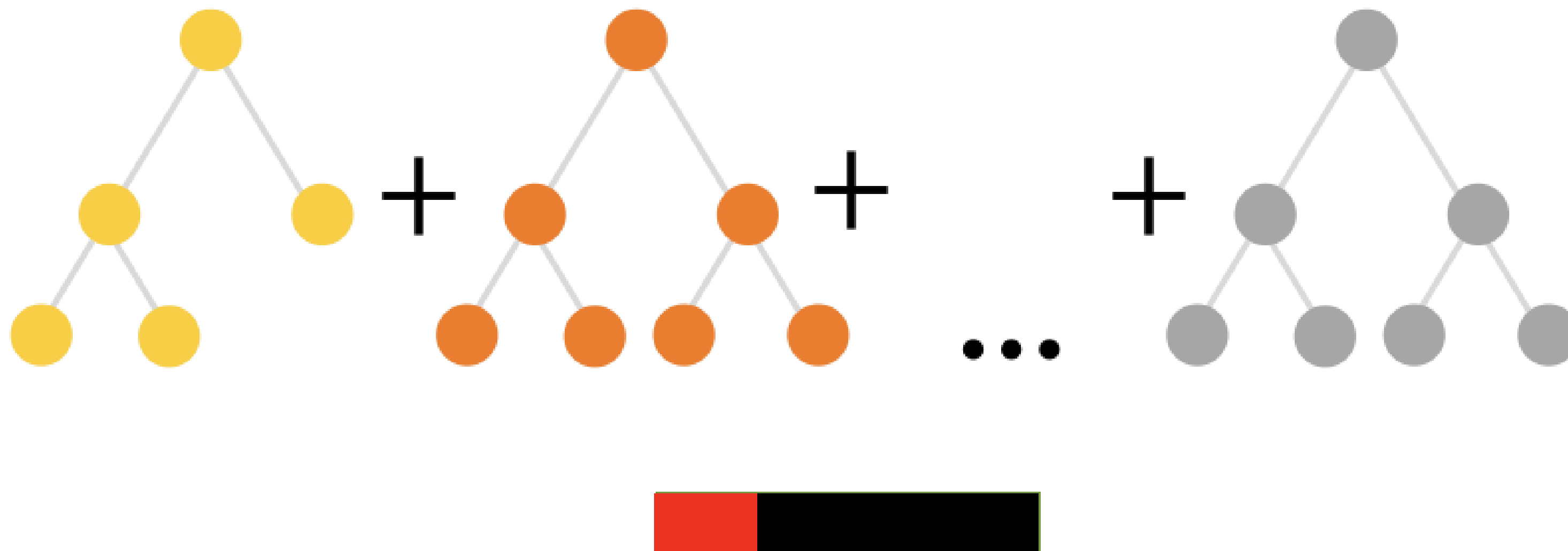
**Недостатки:** слишком много неучтенных факторов, которые «из кабинета не видно»; субъективность специалистов; непрофессиональный эксперт = низкое качество прогнозов.



# Оценка урожайности сельскохозяйственных культур методами машинного обучения

3

**Машинное обучение (*machine learning*, *ML*)** — это наука о разработке алгоритмов и статистических моделей, которые компьютерные системы используют для выполнения задач без явных инструкций, полагаясь на шаблоны и логические выводы.





# Постановка задачи

## Задача

- ✓ Прогнозирование урожайности подсолнечника с помощью методов ML.

## Признаки

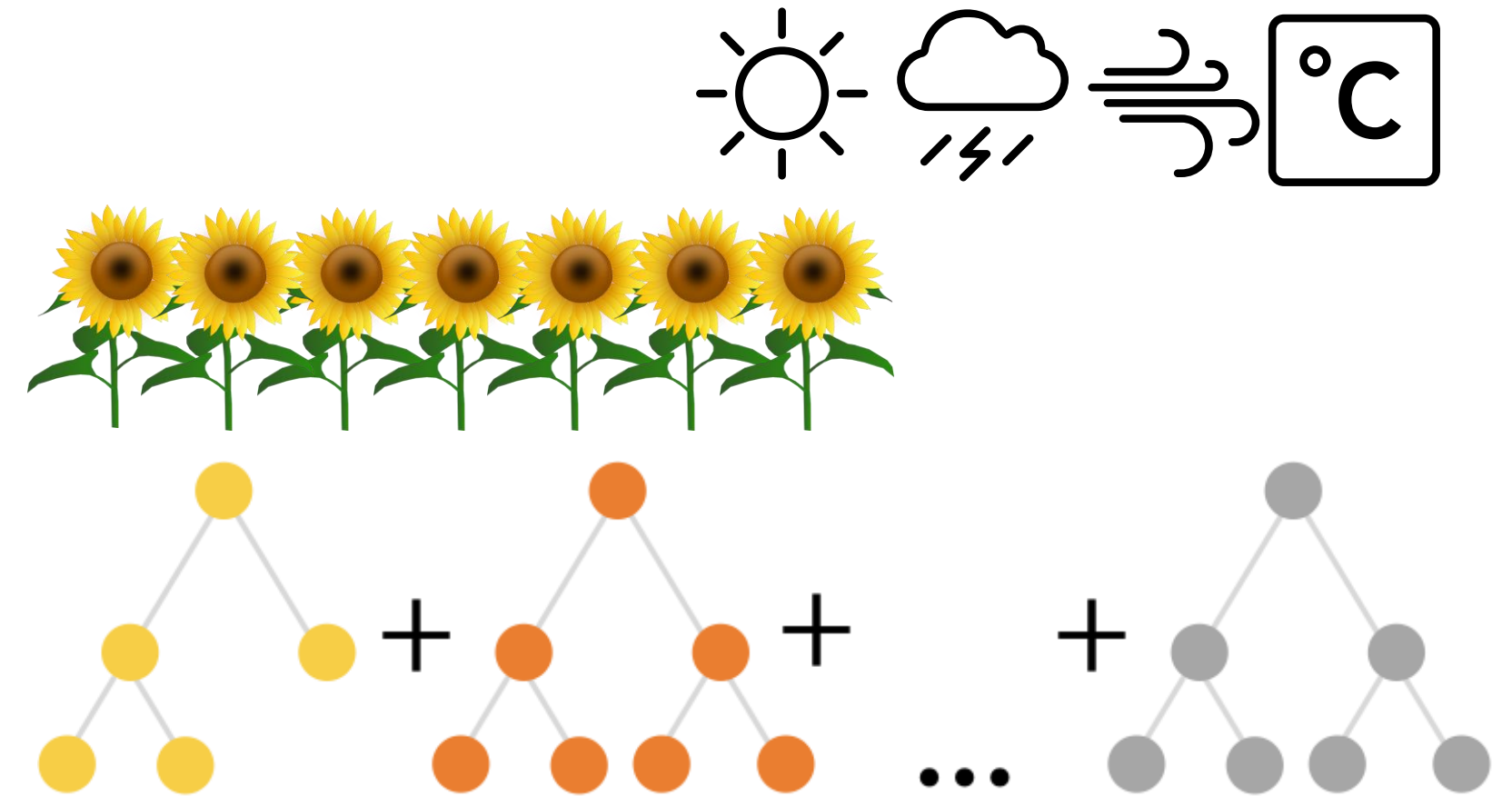
- ✓ Данные гидрометеорологической информации;
- ✓ Паспорт муниципальных районов;
- ✓ Дополнительные авторские фичи.


## Целевая переменная

- ✓ Средняя урожайность подсолнечника по муниципальным районам.

## Метрика качества

- ✓ Root mean square error (RMSE).  
(целевое значение  $RMSE \leq 10\%$ )




$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$



# Входные данные.

2



# Входные данные

# 1 Гидрометеорологическая информация

Источник:

Сайт расписания погоды [pr5.ru](http://pr5.ru)

Набор данных содержит 30 параметров с гидрометеорологической информацией.

## 2 Данные об урожайности подсолнечника

ИСТОЧНИК:

**Официальный сайт Росстат**

Набор данных содержит информацию об урожайности сельскохозяйственных культур по муниципальным районам с 2007 года, всего 11 параметров.

### 3 Информация о населенных пунктах

Источник:

## Сайт платформы ИНИД

Набор данных содержит информацию о населенных пунктах РФ (гео-координаты, административно-территориальная классификация, численность и др.), всего 18 параметров.

Мобильная версия | Главная | Новости | О сайте | Частные вопросы (FAQ) | Контакты | Разместить объявление на рсрп

rp5.ru  
радиосеть погоды
Беларусь Литва Россия Украина Все страны
Название города или села 
🔍 Language 🌐 Единицы измерений ⚙️ Прогноция 📅 Мобильная версия 📱
RSS

---

Все страны • Россия • Москва • Москва (ВДХ)

## Архив погоды в Москве (ВДХ) 📍 См. на карте ✂️ Архив погоды в аэропорту (19 км, -2 °C)

Архив погоды на метеодатчике (2 км, -1,2 °C) ☁️ Прогнозы погоды

номер метеостанции:  , наблюдения с 1 февраля 2005


Смотреть архив погоды
Скачать архив погоды
Статистика погоды

Конечная дата периода: 
Период выборки: ☒ 1 сутки ☐ 7 суток ☐ 30 суток

Для получения пояснений наведите курсор мыши на соответствующий заголовок

Дата / Местное время	T	Po	P	Pa	U	DD	Ft	#10	#3	N	WW	W1	W2	Tn	Tx	Cl	Nh	H	Cm
18	-1.3	745.3	759.4	-0.6	93	Ветер, дульный с востока-северо-востока	Легкий ветер (2 м/c)			100 %	Снег незначительный.	Снег и/или другие виды твердых осадков	Облака покрывали более половины неба в течение всего соответствующего периода.			Слоистые тучнообразные или слоистые разрозненные, либо те и другие, но не относящиеся к облакам сплошной погоды.	100 %	200-300	
15	-1.2	745.9	760.0	0.2	92	Ветер, дульный с северо-северо-востока	Легкий ветер (2 м/c)			100 %			Облака покрывали более половины неба в течение всего соответствующего периода.			Слоистые тучнообразные или слоистые разрозненные, либо те и другие, но не относящиеся к облакам сплошной погоды.	100 %	200-300	
12	-1.1	745.7	759.8	1.3	96	Ветер, дульный с севера	Тихий ветер (1 м/c)			100 %	Дымка.	Облака покрывали более половины неба в течение всего соответствующего периода.	Облака покрывали более половины неба в течение всего соответствующего периода.			Слоистые тучнообразные или слоистые разрозненные, либо те и другие, но не относящиеся к облакам сплошной погоды.	100 %	100-200	
09	-1.2	744.4	758.5	1.3	98	Ветер, дульный с севера	Тихий ветер (1 м/c)			100 %	Дымка.	Облака покрывали более половины неба в течение всего соответствующего периода.	Облака покрывали более половины неба в течение всего соответствующего периода.	-1.2		Слоистые тучнообразные или слоистые разрозненные, либо те и другие, но не относящиеся к облакам сплошной погоды.	100 %	50-100	



2022r



Федеральная служба государственной статистики

[VK](#)
[🔍](#)

Введите свой запрос

Eng

🌐

О Росстате

Статистика

Публикации

Респондентам

Пресс-служба

Контакты

Главная страница / Статистика / Официальная статистика / Базы данных


В ИЗБРАННОЕ

ЕМИСС


Статистический регистр

Базы данных


На Едином Интернет-портале Росстата представлены следующие базы данных

 WEB

Единая межведомственная информационно – статистическая система (ЕМИСС)  
28.08.2018

 HTM



Показатели муниципальных образований  
22.21 Кб, 22.09.2022

 WEB




Статистический регистр Росстата  
27.06.2019


Росстат в соцсетях

Подписывайтесь и следите за деятельностью Федеральной службы государственной статистики в социальных сетях

Поделиться в соцсетях



ИНИД

ИНСТИТУТ НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬНОГО ДАННЫХ

Государственные финансы

Образование и наука

Доходы и неравенство

Здравоохранение

Экономика

Статистика

Фильтр

Режим доступа

☐ Открытый ?

[Сбросить](#)

Орган власти

Поиск

☐ Федеральная служба государственной статистики РОСSTAT

☐ Автономная некоммерческая

Каталог

Исследования

Доступ

О проекте

Новости

Мероприятия

По дате обновления

По скачиваниям

По просмотрам

Показывать по: 10

Открытый доступ

Обновлено: 23.03.2022

**Экспорт и импорт российских регионов: таможенная статистика с детализацией до товаров за 2016–2021 гг.**

Сведения Федеральной таможенной службы по экспорту и импорту. Данные представлены с детализацией до десятизначных кодов товаров, а также стран, в которые (из которых) эти товары импортируются или экспортируются субъектами РФ

6440 471 CSV

Открытый доступ

Обновлено: 27.01.2022

**База данных показателей муниципальных образований России за 2006 – 2020 гг.**

Социально-экономические характеристики муниципальных образований (МО) всех уровней в России за 2006-2020 гг.

19150 1226 CSV POSTGRESQL



# Цели и задачи дипломного проекта

1. Разработка web-скраперов для автоматизации процесса сбора входных данных
2. Создание базы данных для хранения собранной информации на платформе PostgreSQL
3. Разведочный анализ, предобработка, очистка и подготовка данных для обучения и тестирования модели
4. Разработка предиктивной модели для прогнозирования урожайности подсолнечника



# Автоматизация процесса сбора входных данных

## Для автоматизации процесса сбора данных были разработаны web-скраперы

Сбор гидрометеорологической информации с сайта <https://rp5.ru> осуществлялся с помощью web-скрапера из ноутбука «*parser\_rp5.ipynb*». Указанный алгоритм был написан на языке программирования Python с использованием библиотеки selenium.

Мобильная версия | Главная | Новости | О сайте | Частые вопросы (FAQ) | Контакты | Разместить объявление на rp5

БеларусьЛитваРоссияУкраинаВсе страны

Название города или села

Language

Единицы измерений

Приложения

Мобильная версия

RSS

Все страныРоссияМоскваМосква (ВДНХ)

Архив погоды в Москве (ВДНХ)См. на картеАрхив погоды в аэропорту ( 19 км, -2 °C )

Архив погоды на метеодатчике ( 2 км, -1.2 °C )Прогноз погоды

номер метеостанции 27612, наблюдения с 1 февраля 2005

Смотреть архив погодыСкачать архив погодыСтатистика погоды

Конечная дата периода: 11.12.2022Период выборки: 1 сутки7 суток30 суток

Для получения пояснений наведите курсор мыши на соответствующий заголовок

Дата / Местное время	T	Po	P	Pa	U	DD	Ff	ff10	ff3	N	WW	W1	W2	Tn	Tx	CI	Nh	H	Cm
18	-1.3	745.3	759.4	-0.6	93	Ветер, дующий с востоко-северо-востока	Легкий ветер (2 м/с)			100 %	Снег нелинейный.	Снег и/или другие виды твердых осадков	Облака покрывали более половины неба в течение всего соответствующего периода.			Слоистые туманообразные или слоистые разорванные, либо те и другие, но не относящиеся к облакам плохой погоды.	100 %	200-300	
15	-1.2	745.9	760.0	0.2	92	Ветер, дующий с северо-северо-востока	Легкий ветер (2 м/с)			100 %						Слоистые туманообразные или слоистые разорванные, либо те и другие, но не относящиеся к облакам плохой погоды.	100 %	200-300	
12	-1.1	745.7	759.8	1.3	96	Ветер, дующий с севера	Тихий ветер (1 м/с)			100 %	Дымка.	Облака покрывали более половины неба в течение всего соответствующего периода.	Облака покрывали более половины неба в течение всего соответствующего периода.			Слоистые туманообразные или слоистые разорванные, либо те и другие, но не относящиеся к облакам плохой погоды.	100 %	100-200	
09	-1.2	744.4	758.5	1.3	98	Ветер, дующий с севера	Тихий ветер (1 м/с)			100 %	Дымка.	Облака покрывали более половины неба в течение всего соответствующего периода.	Облака покрывали более половины неба в течение всего соответствующего периода.	-1.2		Слоистые туманообразные или слоистые разорванные, либо те и другие, но не относящиеся к облакам плохой погоды.	100 %	50-100	

2022г.  
11 декабря

Сбор информации по урожайности сельхозкультур с сайта Росстат осуществлялся с помощью web-скрапера из ноутбука «*parser\_rosstat.ipynb*». Данный web-скрапер также был написан на языке Python с использованием библиотеки selenium.

Федеральная служба государственной статистики

Введите свой запрос

Eng

О РосстатеСтатистикаПубликацииРеспондентамПресс-службаКонтакты

Главная страница / Статистика / Официальная статистика / Базы данныхВ ИЗБРАННОЕ

ЕМИСС

Статистический регистр

Базы данных

На Едином Интернет-портале Росстата представлены следующие базы данных

WEB

Единая межведомственная информационно – статистическая система (ЕМИСС)

28.08.2018

HTM

Показатели муниципальных образований

22.21 Кб, 22.09.2022

WEB

Статистический регистр Росстата

27.06.2019

Росстат в соцсетях

Подписывайтесь и следите за деятельностью Федеральной службы государственной статистики в социальных сетях

ВКонтакте

Одноклассники

Поделиться в соцсетях



# Организация процесса хранения данных

Для хранения входной информации была создана реляционная БД «*weather*» (на open-source платформе **PostgreSQL**). Для создания базы и индексирования таблиц использовался код из файла «*weather\_DB.sql*».

В базе созданы три таблицы: «*weather*», «*settlement*», «*yield*» Для выгрузки данных из базы, был использован код из файла «*sample\_for\_ml\_model.sql*».

weather x

Свойства

Диаграмма

Название: weather

Namespace ID: 17045

Комментарий:

Владелец: postgres

	Название	ID объекта	Владелец	Табличное пространство	Примерное число строк
Таблицы	settlement	25 358	postgres	pg_default	155 922
	weather	25 277	postgres	pg_default	67 012 648
	yield	25 377	postgres	pg_default	22 562

Представления

Мат. представления

Индексы

Функции

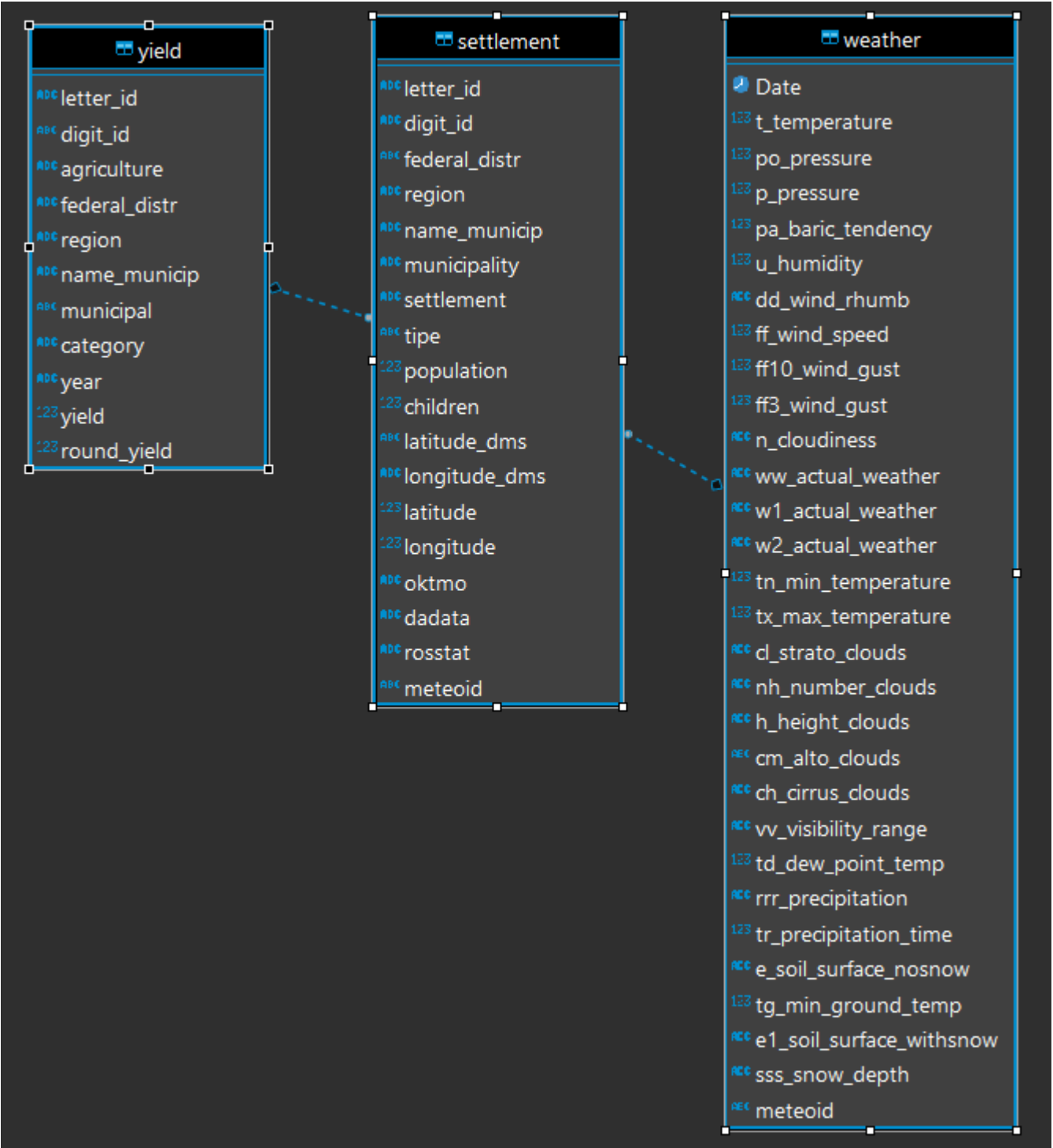
Последовательности

Типы данных

Агрегатные функции

Права доступа

Исходный код



Связи между таблицами  
«*weather*» - «*settlement*» foreign key «*meteoid*»  
«*yield*» - «*settlement*» foreign key «*digit\_id*»





# Разведочный анализ и предобработка данных.



3



# Цели и задачи дипломного проекта

1. Разработка web-скраперов для автоматизации процесса сбора входных данных
2. Создание базы данных для хранения собранной информации на платформе PostgreSQL
3. Разведочный анализ, предобработка, очистка и подготовка данных для обучения и тестирования модели
4. Разработка предиктивной модели для прогнозирования урожайности подсолнечника



# Выгрузка из БД

1. Данные из таблица «**weather**» были сгруппированы по неделям.
2. Данные с гидрометеорологической информацией были разделены на группы по показателям:
  - a. температура (воздуха, почвы, точки росы);
  - b. атмосферное давление;
  - c. влажность воздуха;
  - d. ветер (направление, скорость);
  - e. облачность (высота, форма облаков);
  - f. горизонтальная видимость;
  - g. количество выпавших осадков;
  - h. состояние поверхности почвы.
3. Было произведено агрегирование каждого показателя по min, max, average, sum.
4. Данные из таблиц «**settlement**» и «**yield**» были выгружены без изменений.
5. Описание содержания таблиц БД «**weather**» представлено в файле «**description\_project.txt**»





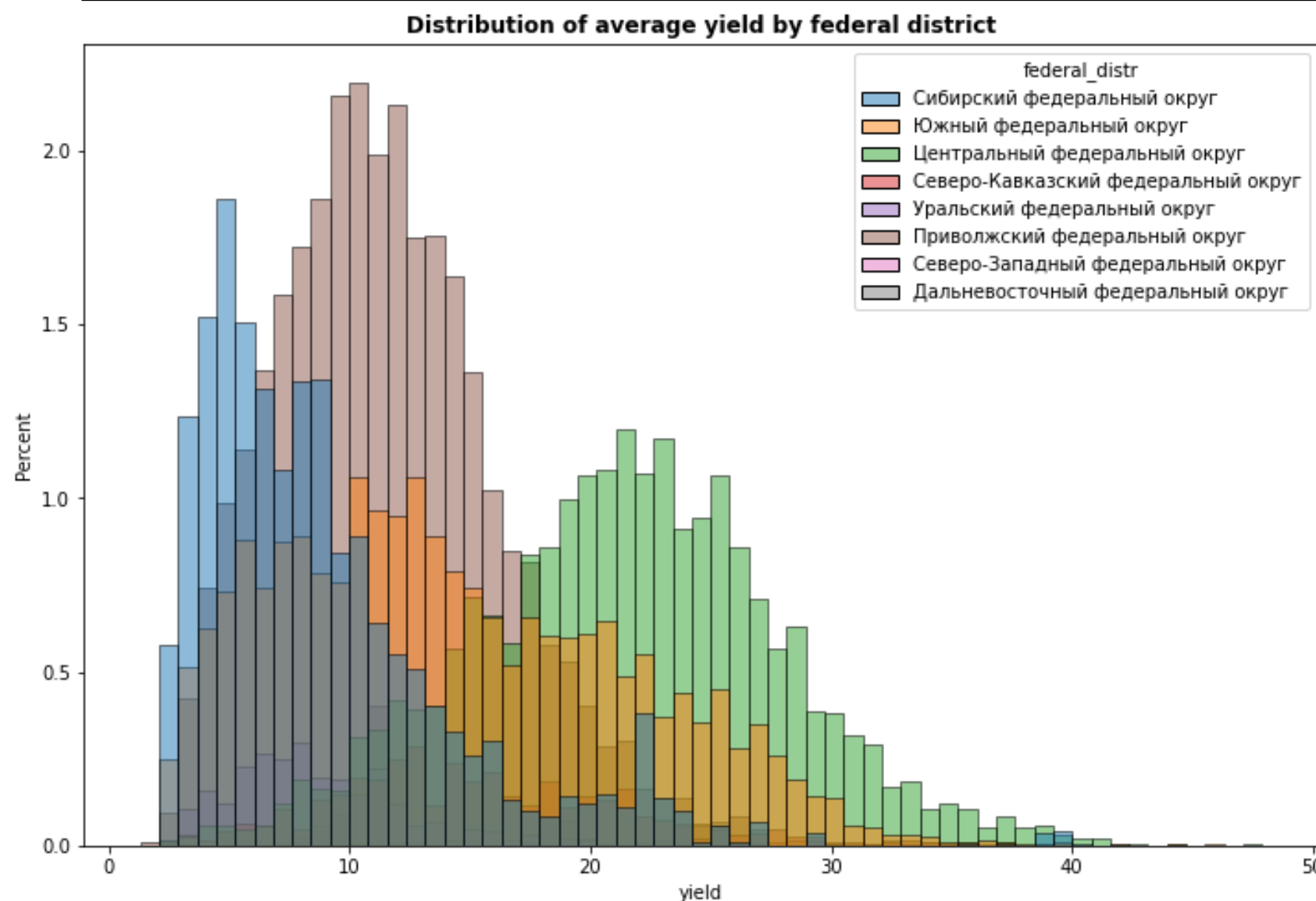
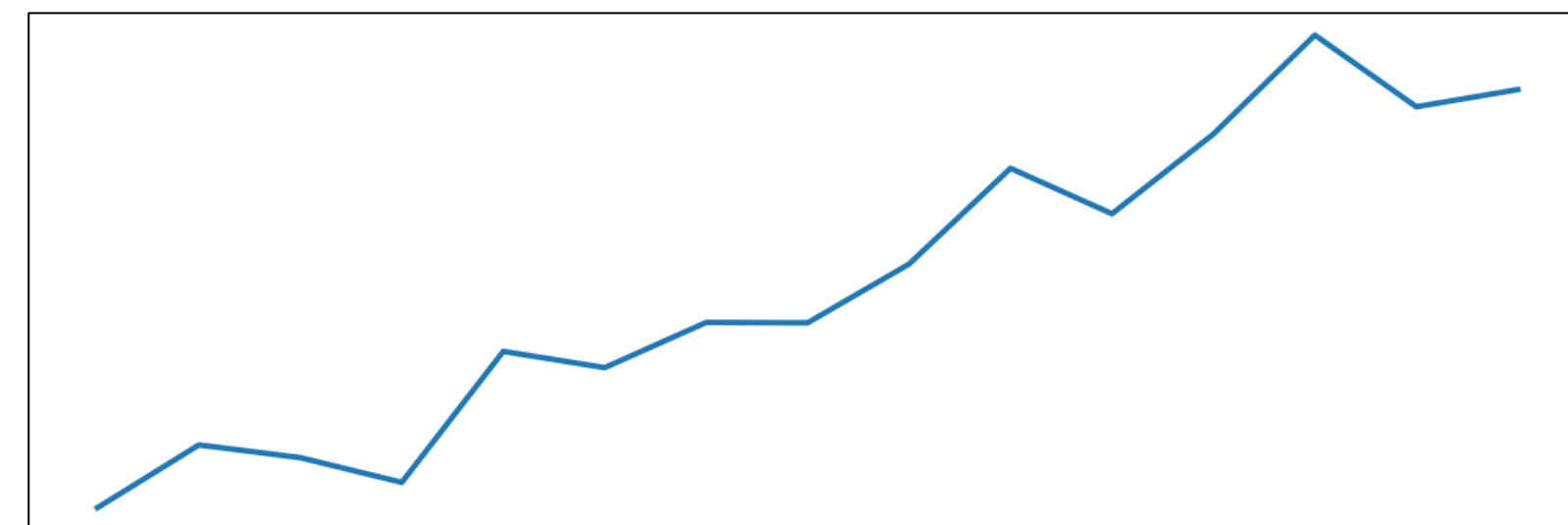
# Разведочный анализ параметров гидрометеорологической информации

1. Каких-либо значимых аномалий или отклонений в числовых данных нет.
2. Некоторые данные имеют пропущенные значения.
3. Признаки из одной группы имеют очень высокую степень корреляции. В целом, корреляция переменных не очень высокая.
4. Для прогнозирования целевой переменной планируется использовать модели градиентного бустинга. Поэтому некоторые признаки, имеющие высокую корреляцию, было решено оставить.
5. Распределение значений переменных, агрегированных по средней, стремится к форме нормального распределения.

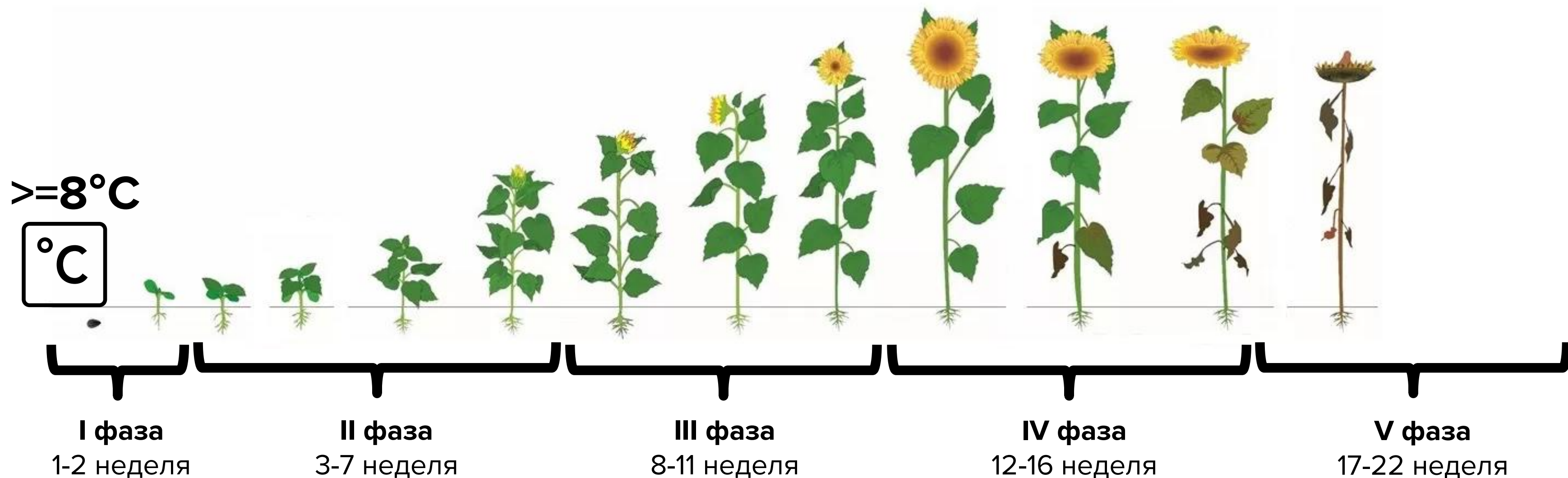


# Разведочный анализ целевой переменной

1. Территориальное расположение муниципального района – один из важных признаков при определении величины урожайности.
2. Динамика урожайности по годам имеет трендовую составляющую. Для выявления трендовой компоненты планируется использовать SSA.
3. Для повышения эффективности прогнозирования урожайности, планируется провести обучение моделей на данных по каждому региону отдельно.
4. Распределение объектов по регионам неравномерное. Обучение моделей будет проводиться на наборах данных, имеющих более 30 объектов.



# Схема цикла вегетации подсолнечника



Оптимальные сроки сева подсолнечника, наступают когда верхний слой почвы прогреется до  $8^{\circ}\text{C}$  и выше.

Средний срок вегетации подсолнечника – 150-155 дней (сильно варьируется в зависимости от сортов и гибридов – 90-160 дней).

Условно, выделяют 5 фаз роста и развития растений (всходы, бутонизация, цветение, созревание, хозяйственная спелость).





# Трансформация данных для создания тренировочных и тестовых датасетов

I. Для каждого уникального *meteoid* и *year* создаем мини-датасет, где мин. индекс (*i\_start*) начинается с недели на которой *t° почвы*  $\geq 8\text{ }^{\circ}\text{C}$ , а макс. индекс (*i\_end*) равен *i\_start*+21

Исходные данные

Date	Year	Features	meteoid
week1	2007	temp., press, etc	1
...	...	(min, max, avg, sum)	1
week52	2007		1
...	...	...	...
week1	2021	temp., press, etc	n
...	...	(min, max, avg, sum)	n
week52	2021		n

II. Все features в датасете группируются и агрегируются по соответствующим фазам (по номерам соотв. недель)

Создание набора данных по сезонам

Date	Year	Features	meteoid	Phase
weeks for I phase ( $t \geq 8^{\circ}\text{C}$ )	2007	agg. feat.	1	1
weeks for II phase	2007	agg. feat	1	2
weeks for III phase	2007	agg. feat	1	3
weeks for IV phase	2007	agg. feat	1	4

- 1
- 2
- 3
- 4

III. Строки мини-датасета транспонируются в вектор (строки с уникальными phase конкатенируются)

Выходной датасет

1

+

2

+

3

+

4

+

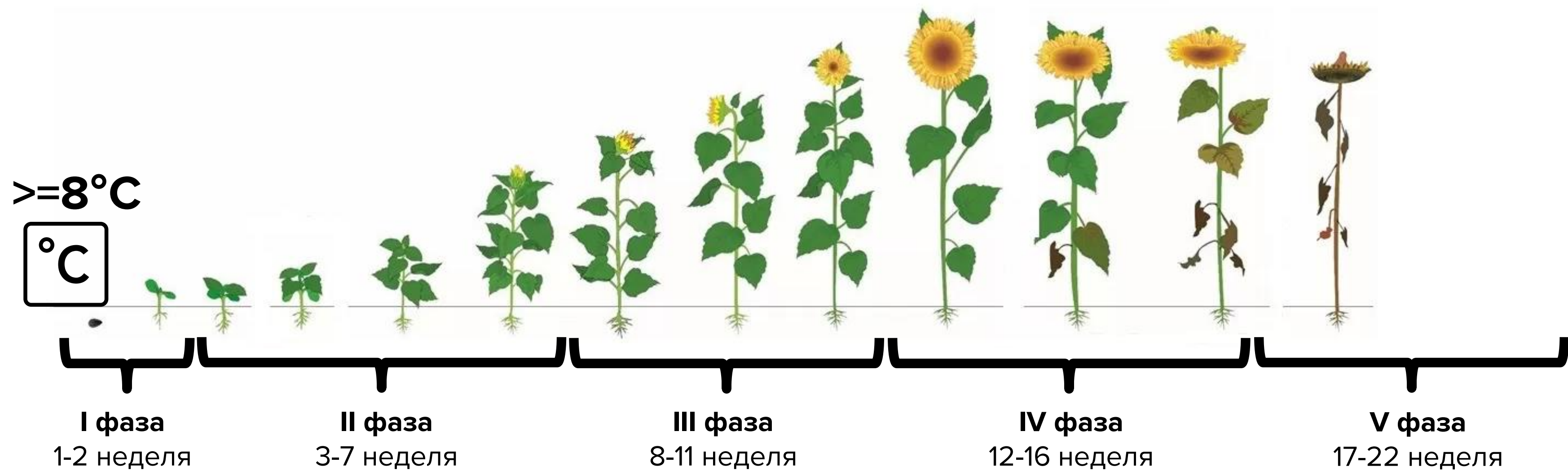
tot

Year	meteoid	Features
uniq.	uniq.	data for phase I-IV and whole season

IV. Добавляем к вектору строку с агрегированными данными за весь сезон (за I-IV фазы)



# Входной вектор признаков для обучения модели



Year	meteoid	Features					= target
unique	unique	features for phase I	features for phase II	features for phase III	features for phase IV	features for whole season	

Для обучения *на вход в модель подается вектор признаков*, который содержит одинаковые переменные, сгруппированные по фазам вегетации, плюс признаки агрегированные за весь сезон.

*Целевая переменная* – урожайность подсолнечника по муниципальному району за определенный год.



# Виды датасетов для обучения модели



Year	meteoid	Features				
unique	unique	features for phase I	features for phase II	features for phase III	features for phase IV	features for whole season

«phase»  
(165 feat.)

Year	meteoid	Features
unique	unique	features for whole season

«total»  
(45 feat.)

С цель сокращения использования вычислительных ресурсов, оптимизации моделей и улучшения их интерпретации был также создан датасет с сокращенным количеством переменных (датасет типа «total» - 45 переменных). В этом датасете признаки агрегировались по всему сезону (см. набор данных «*total\_data*» и «weather» из ноутбука «*preparation\_mldata\_4phase.ipynb*»). Датасет «phase» содержит 165 переменных.





# Добавление новых признаков / удаление малоинформативных признаков

Были добавлены следующие features:

1. **Гидротермический коэффициент Селянинова** (ГТК) –  
feature name GTK

$$GTK = \frac{(\sum precipitation) \times 10}{\sum air\ temperature}$$

2. **Суховей** (ОПЯ\* суховей) – feature name dry\_wind

$$dry\_wind = \frac{avg.\ wind\ speed \times avg.\ air\ temperature}{avg.\ humidity}$$

3. **Индикатор переувлажнения** – feature name  
precipitation\_speed

$$precipitation\_speed = \frac{(\sum precipitation)}{\sum precipitation\ time}$$

4. **Диапазон фичи** (разница max-min) – features name  
diff\_<feature name>

$$diff\_ < feat.name > = \\ max\_ < feat.name > - min\_ < feat.name >$$

5. **Границы района** – max, min и среднее значение гео-  
координат (долготы и широты) по муницип. району,  
features <...>\_border, central\_lat, central\_lon

6. **Количество населенных пунктов** в районе.

7. **Севооборот** – значение каждого 4-го года (от 0 - до 3)

**Были удалены следующие features:** признаки, которые имели низкие веса в листе feature importance по результатам обучения моделей CatBoost и XGBoost (на параметрах default)

\* ОПЯ – опасное природное явление





# Предиктивная модель для прогнозирования урожайности.

4



# Цели и задачи дипломного проекта

1. Разработка web-скраперов для автоматизации процесса сбора входных данных
2. Создание базы данных для хранения собранной информации на платформе PostgreSQL
3. Разведочный анализ, предобработка, очистка и подготовка данных для обучения и тестирования модели
4. **Разработка предиктивной модели для прогнозирования урожайности подсолнечника**



# Типы обучаемых моделей

## Модели регрессии

### I. Модели градиентного бустинга:

- CatBoost Regressor;
- XGB Regressor.

### II. Модель линейной регрессии

- Sklearn Linear Regression

## Библиотеки



CatBoost

*XGBoost*



Keras

## Нейронные сети

### I. LSTM Bidirectional.

### II. Полносвязная нейронная сеть



# Результат обучения LSTM Bidirectional

## Результаты обучения модели.

Train Score: 1.23 RMSE  
Test Score: **3.49 RMSE**

## Комментарий к модели:

- Обучение и тестирование модели проводилось на датасете с сокращенным количеством переменных («total») – 45 признаков;
- Размер batch – 32;
- Количество эпох – 1000
- Наименование файла с результатами «lstm\_tot\_at\_epoch\_{epoch}.h5»

## LSTM Bidirectional Summary.

Model: "LSTM Bidirectional"		
Layer (type)	Output Shape	Param #
=====		
lstm (LSTM)	(None, 1, 42)	14280
<hr/>		
lstm_1 (LSTM)	(None, 1, 84)	42672
<hr/>		
bidirectional (Bidirectional	(None, 1, 168)	113568
<hr/>		
lstm_4 (LSTM)	(None, 1, 42)	35448
<hr/>		
lstm_5 (LSTM)	(None, 10)	2120
<hr/>		
dense (Dense)	(None, 1)	11
=====		
Total params: 208,099		
Trainable params: 208,099		
Non-trainable params: 0		
<hr/>		





# Результат обучения полносвязной нейронной сети

## Результаты обучения модели.

Train Score: 5.97 RMSE  
Test Score: **6.48 RMSE**

## Комментарий к модели:

- Обучение и тестирование модели проводилось на датасете с сокращенным количеством переменных («total») – 45 признаков;
- Размер batch – 32;
- Количество эпох – 1000
- Наименование файла с результатами «nnmodel\_at\_epoch\_{epoch}.h5»

## Полносвязная нейронная сеть Summary.

Model: "Sequential"		
Layer (type)	Output Shape	Param #
=====		
dense_1 (Dense)	(None, 1, 42)	1806
batch_normalization (BatchNo	(None, 1, 42)	168
dense_2 (Dense)	(None, 1, 42)	1806
dropout (Dropout)	(None, 1, 42)	0
batch_normalization_1 (Batch	(None, 1, 42)	168
dense_3 (Dense)	(None, 1, 20)	860
dropout_1 (Dropout)	(None, 1, 20)	0
batch_normalization_2 (Batch	(None, 1, 20)	80
dense_4 (Dense)	(None, 1, 10)	210
dense_5 (Dense)	(None, 1, 1)	11
=====		
Total params: 5,109		
Trainable params: 4,901		
Non-trainable params: 208		
=====		



# Результаты обучения нейронных сетей

Результаты обучения нейронных сетей.

Model name	RMSE		test to train
	train	test	
LSTM Bidirectional	1.23	3.49	+2.26
Fully connected NN	5.97	6.48	+0.51
<b>LSTM vs. Full. conn. NN</b>	<b>-4.74</b>	<b>-2.99</b>	

## Комментарий:

- Обучение и тестирование нейронных сетей проводилось на датасете с сокращенным количеством переменных («total») – 45 признаков;
- Рекуррентная нейросеть показала более лучшие результаты, как на тренировочном, так и на тестовом датасете.



# Результаты обучения регрессионных моделей на default - параметрах

Результаты обучения регрессионных моделей.

Model name	RMSE test		phase to total
	total	phase	
Linear Regression	114 676.876	-	-
CatBoost Regressor	2.811	2.747	-0.064
XGB Regressor	2.147	2.119	-0.028
Best results	2.147	2.119	-0.028

## Комментарий:

- Обучение и тестирование модели проводилось на датасете с полным и сокращенным количеством переменных («phase» и «total») – 165 и 45 признаков;
- Модели показывают более лучшие результаты на полном датасете («phase»);
- Более лучшие результаты на обоих датасетах показала модель XGB Regressor;
- Модель Linear Regression показала худшие результаты ( $r^2 = 0.2$ ).



# Подбор оптимальных параметров модели с помощью метода `grid_search`



Оптимизированные параметры  
регрессионных моделей



CatBoost

**XGBoost regressor:**

```
{
    'colsample_bytree': 0.6,
    'gamma': 0.05,
    'learning_rate': 0.075,
    'max_depth': 10,
    'min_child_weight': 1,
    'n_estimators': 5000,
    'objective': 'reg:squarederror',
    'random_state': 2,
    'subsample': 1
}
```

**CatBoost regressor:**

```
{
    'loss_function': 'RMSE',
    'verbose': False,
    'max_leaves': 64,
    'depth': 6,
    'random_seed': 2,
    'iterations': 5000,
    'learning_rate': 0.075,
    'l2_leaf_reg': 0.07
}
```





# Результаты обучения моделей с оптимизированными параметрами

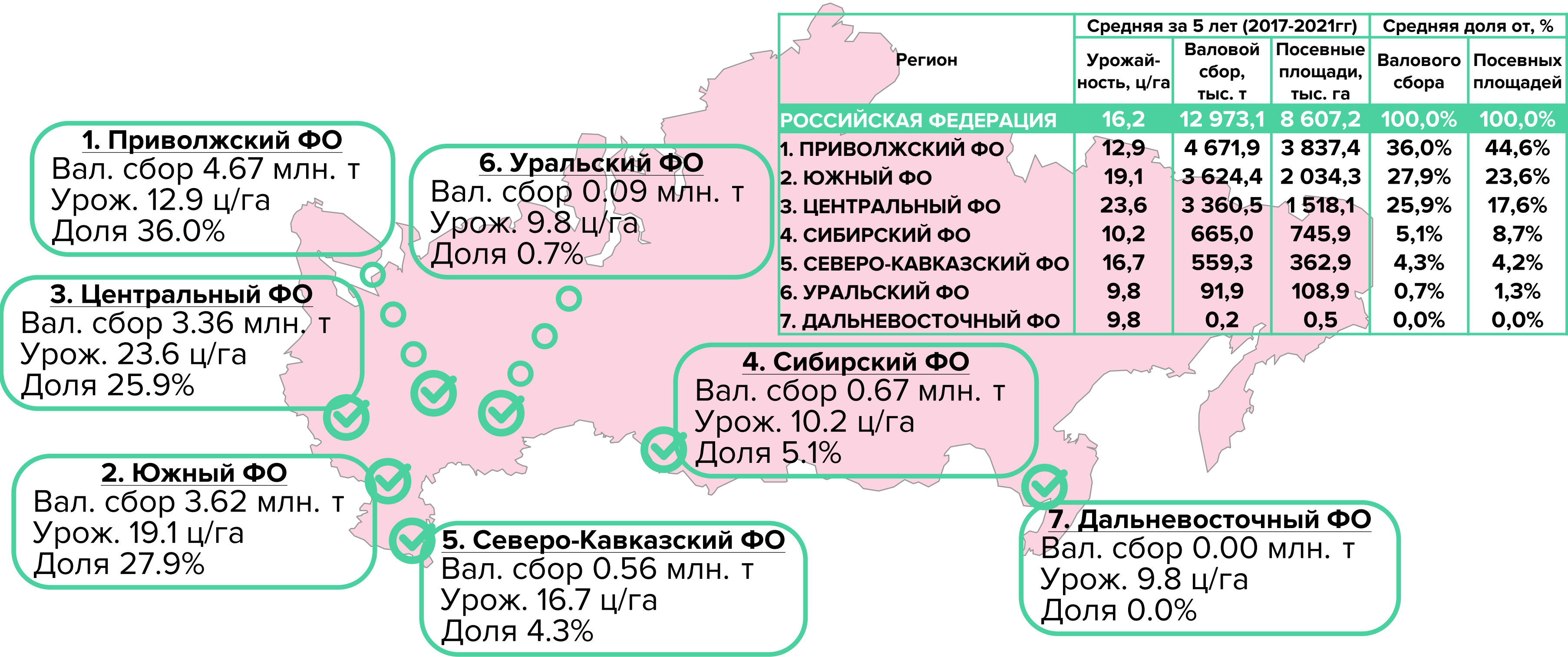
Model name	RMSE test		phase to total	RMSE test	
	total	phase		total	phase
На общих данных (general)				optimal to default	
CatBoost Regressor	2.635	2.164	-0.471	-0.176	-0.583
XGB Regressor	2.022	1.979	-0.043	0.125	-0.140
Best results	2.022	1.979	-0.043		

## Комментарий:

- Результаты обучения и тестирования моделей с оптимизированными параметрами лучше, чем с default-параметрами;
- Результаты обучения моделей на датасете с полным количеством признаков («phase») лучше, чем на сокращенном («total»);



# География регионов производства семян подсолнечника в России



# Результаты обучения моделей с оптимизированными параметрами

- ✓ 1. Проведено обучение и тестирование моделей на данных по 40 регионам
- ✓ 2. Доля производства семян подсолнечника в указанных регионах – составляет 99,8%
- ✓ 3. Совокупная доля посевных площадей данных регионов – составляет 99,7%
- ✓ 4. Взвешенное значение RMSE по регионам дает лучшее значение, чем простое среднее:
  - CatBoost reg. – значение  $RMSE_{wg} = 1.498$  (vs.  $RMSE = 2.273$ )
  - XGBoost reg. – значение  $RMSE_{wg} = 1.601$  (vs.  $RMSE = 2.273$ )
- ✓ 5. Использование лучшего значения RMSE по моделям снижает до  $RMSE_{wg} = 1.458$

$$RMSE_{wg} = \sum (w_{reg} \times RMSE_{reg})$$

где:  $RMSE_{wg}$  – взвешенное значение RMSE;

$w_{reg}$  – доля региона в общем объеме производства;

$RMSE_{reg}$  – значение RMSE модели для региона.



# Общие результаты обучения регрессионных моделей

Регион	Средняя за 5 лет		CatBoost regressor			XGBoost regressor			Best RMSE value of models		
	Урожай-ность	Доля произ-ва	RMSE	RMSE <sub>wg</sub>	RMSE <sub>wg</sub> to RMSE	RMSE	RMSE <sub>wg</sub>	RMSE <sub>wg</sub> to RMSE	RMSE	RMSE <sub>wg</sub>	RMSE <sub>wg</sub> to RMSE
РОССИЙСКАЯ ФЕДЕРАЦИЯ	16,2	100,0%	2,273	1,498	-0,775	2,273	1,601	-0,672	2,118	1,458	-0,660
1. ПРИВОЛЖСКИЙ ФО	12,9	36,0%	2,131	1,283	-0,848	2,208	1,447	-0,761	1,945	1,198	-0,747
2. ЮЖНЫЙ ФО	19,1	27,9%	1,990	2,117	0,126	2,030	2,179	0,150	1,981	2,115	0,134
3. ЦЕНТРАЛЬНЫЙ ФО	23,6	25,9%	1,732	1,225	-0,507	1,677	1,296	-0,381	1,677	1,203	-0,475
4. СИБИРСКИЙ ФО	10,2	5,1%	2,477	1,071	-1,407	2,593	1,259	-1,334	2,477	1,071	-1,407
5. СЕВЕРО-КАВКАЗСКИЙ ФО	16,7	4,3%	3,487	1,392	-2,096	3,562	1,334	-2,228	3,293	1,319	-1,974
6. УРАЛЬСКИЙ ФО	9,8	0,7%	1,829	1,706	-0,123	1,918	1,795	-0,124	1,829	1,706	-0,123
7. ДАЛЬНЕВОСТОЧНЫЙ ФО	9,8	0,0%	2,357	2,201	-0,156	1,273	0,488	-0,785	1,273	0,488	-0,785





# Итоговые результаты модели прогнозирования урожайности подсолнечника

Результаты обучения регрессионных моделей.

Параметр	Значение	В % к средней урожайности
Средняя урожайность, ц/га	16,2	-
Значение $RMSE_{wg}$	1,458	9,0%
CatBoost	1,498	9,3%
XGBoost	1,601	9,9%



# Прогнозирование урожайности подсолнечника по данным гидрометеорологической информации

**Олег Воропаев**  
Отраслевой аналитик, к.э.н.



olegovoropaev@mail.ru

