ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к дипломному проекту на тему: «Прогнозирование урожайности сельскохозяйственных культур (подсолнечника) по данным гидрометеорологической информации»

Целью данного дипломного проекта является разработка предиктивной модели для прогнозирования урожайности подсолнечника по данным гидрометеорологической информации.

В сельскохозяйственном производстве, как и в любой другой отрасли, эффективное прогнозирование является одним из главных конкурентных преимуществ организации. Точный и своевременный прогноз определяет стратегию работы компании и повышает эффективность реализуемых мероприятий.

Отличительной особенностью отрасли растениеводства в России является получение продукции (валового сбора) один раз в сезон. Начало и окончание сельскохозяйственного сезона не совпадает с календарным годом (см. рис. 1). Так, началом сезона, для производства подсолнечника, является старт уборки урожая текущего года (начало: сентябрь текущего года), а окончанием сезона, период до начала уборки урожая следующего года (окончание: август следующего года).

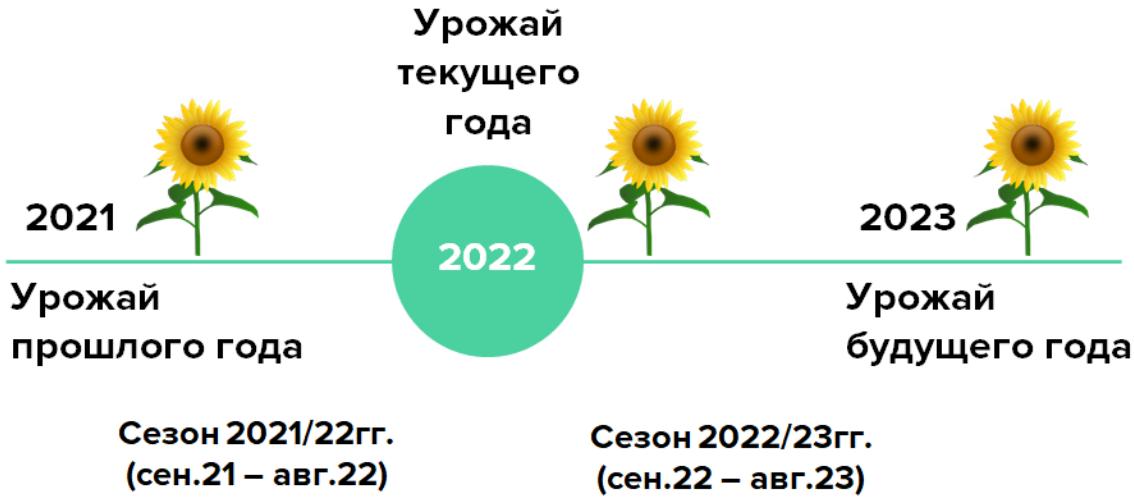


Рисунок 1. Схема сезонности производства подсолнечника в РФ

Одной из главных задач отраслевого аналитика (crop-аналитика) является своевременное и точное прогнозирование валового сбора перед началом нового сезона.

Объем валового сбора сельскохозяйственной культуры оказывает сильное влияние на динамику цен и задает направление тренда в течение всего сезона. В случае сильного снижения валового сбора – на рынке возникает дефицит продукции, что может вызвать рост цен на сырье, а в случае сильного увеличения – появляется избыток предложения и это может оказывать давление на цены.

Объем валового сбора – это производная от двух величин: размера посевной площади и урожайности сельхозкультуры:

$$ВС = ПП \times У,$$

где: ВС – валовой сбор (центнеры, тонны и т.п.)

ПП – размер посевных площадей (гектары (га))

У – урожайность сельскохозяйственной культуры (тонн с га, центнер с га и т.д.)

Старт массовой посевной кампании подсолнечника в РФ начинается в середине апреля, а завершается в конце июня (сроки начала посевной зависят от погодных условий и могут варьироваться в зависимости от региона). Как правило, к 1 июля у аналитиков уже

есть представление о размерах посевных площадей культуры. Таким образом, к началу нового сезона неизвестной переменной остается показатель ***урожайности***.

К концу августа – началу сентября большая часть растений достигает фазы «восковой спелости», что позволяет проводить предварительную оценку урожайности традиционными методами. К традиционным и самым распространенным методам оценки урожайности сельскохозяйственных культур относят: ***кроп-тур*** и ***экспертную оценку***.

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). К ***преимуществам*** данного метода можно отнести: профессиональную оценку урожайности, так как оценка проводится профильными специалистами отрасли; при выезде в поля можно увидеть то, что «не видно из кабинета», оценить реальную обстановку в полях; широкий охват различных регионов (если маршрут кроп-тура продолжительный). К ***недостаткам***: высокие финансовые затраты на организацию поездки (оплата работ, оборудования, транспорта); необходимость точно угадать со сроками поездки (пораньше-плохо, попозже еще хуже); субъективность специалистов.

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 во многом зависят от опыта и знаний эксперта. ***Преимущества*** метода: 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.

В данном дипломном проекте предлагается реализовать решение задачи оценки урожайности сельскохозяйственных культур с помощью методов машинного обучения[1]. Постановка задачи будет звучать следующим образом:

Задача:

- Прогнозирование урожайности подсолнечника с помощью методов ML.

Признаки:

- Данные гидрометеорологической информации;
- Паспорт муниципальных районов;
- Дополнительные добавленные признаки.

Целевая переменная:

- Средняя урожайность подсолнечника по муниципальным районам.

Метрика качества:

- Среднеквадратичная ошибка (RMSE), целевое значение ≤ 10%:

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

В проекте были использованы следующие входные данные (все данные были получены из открытых источников):

1. Статистика по урожайности подсолнечника по муниципальным районам субъектов РФ (источник: Росстат, https://rosstat.gov.ru/). Набор данных содержит информацию об урожайности сельскохозяйственных культур по муниципальным районам с 2007 года, всего 11 параметров.
2. Данные по населенным пунктам РФ с координатами геолокации (для объединения данных по муниципальным районам и метеостанций, источник: ИНИД – Инфраструктура

[1] ***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.

научно-исследовательских данных, https://www.data-in.ru/). Набор данных содержит информацию о населенных пункта РФ (гео-координаты, административно-территориальную классификацию, численность населения и др.), всего 18 параметров.

3. Архив данных гидрометеорологических параметров с наземных метеостанций РФ (источник: https://rp5.ru/). Набор данных содержит 30 параметров с гидрометеорологической информацией.

Для автоматизации процесса сбора исходных данных были написаны web-скраперы. Сбор гидрометеорологической информации с сайта https://rp5.ru осуществлялся с помощью web-скрапера из ноутбука «***parser_rp5.ipynb***», данных по урожайности сельхозкультур с сайта Росстат осуществлялся с помощью скрапера из ноутбука «***parser_rosstat.ipynb***». Указанные алгоритмы были написаны на языке программирования Python 3 с использованием библиотеки selenium. Целесообразность использования автоматизированного сбора данных была продиктована объективной необходимостью[2], с целью экономии времени и трудовых ресурсов.

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***».

В ходе предварительной обработки и анализа данных был проведен разведочный анализ данных, по результатам которого были сделаны следующие выводы:

1. Каких-либо значимых аномалий или отклонений в числовых данных нет.
2. Некоторые данные имеют пропущенные значения, которые в процессе были заполнены либо средними многолетними значениями по неделям, либо через зависимости от других признаков.
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.

Был также проведен разведочный анализ данных целевой переменной, выводы следующие:

1. Территориальное расположение муниципального района является одним из важнейших признаков при определении величины урожайности.
2. Динамика урожайности по годам имеет трендовую составляющую. Здесь возможно влияние других факторов, независящих от погоды, таких как совершенствование технологии возделывания культур, использование более лучшего посевного материала (результат современных достижений селекции и генетики) и др. Трендовая компонента должна быть выявлена и удалена. Так как, наличие других факторов может ухудшить выявление взаимосвязей влияния от факторов погоды. 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 33 объектов (30 на обучение, 10% на тест).

---

[2] Так, например, гидрометеорологическая информация была собрана с 1685 наземных метеостанций расположенных на территории РФ. Архив данных по некоторым метеостанциям содержал информацию начиная с 2005 года, на сайте существуют ограничения связанные с объемом информации, который может быть скачан за один раз (максимум можно скачать архив за 5 лет). Таким образом при ручном скачивании всего вышеописанного объема потребовалось бы совершить не менее 6,5 тыс. итераций.

Для более эффективной работы модели была проведена трансформация входных данных в вектор признаков отражающий влияние гидрометеорологических факторов на растения в период их роста и развития (см. ноутбук «***preparation_mldata_4phase.ipynb***»).

Преобразование данных строилось с учетом природного цикла вегетации подсолнечника (см. рис. 2). Для выращивания данной культуры необходимо соблюдение ряда агротехнологических требований. Так, оптимальные сроки сева подсолнечника, наступают когда верхний слой почвы прогреется до 8°С и выше. Средний срок вегетации подсолнечника – 150-155 дней (сильно варьируется в зависимости от сортов и гибридов – 90-160 дней). В период вегетации, условно, выделяют 5 фаз роста и развития растений (всходы, бутонизация, цветение, созревание, хозяйственная спелость).

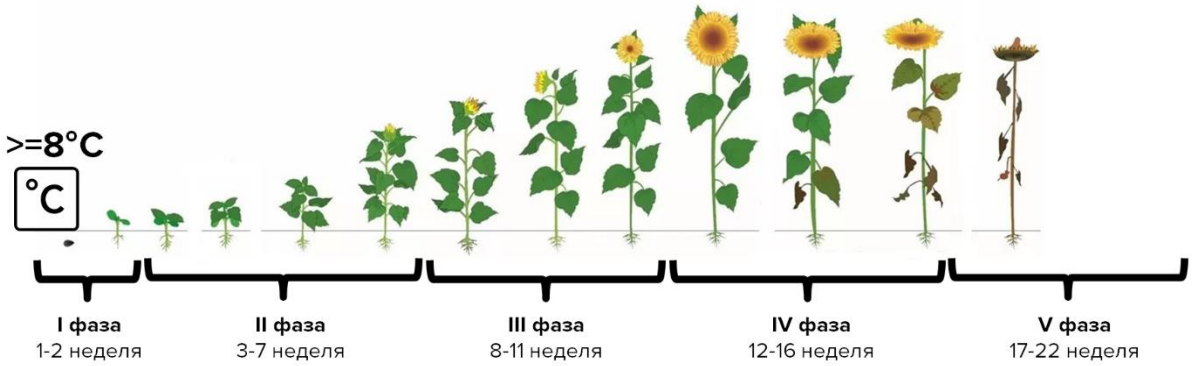


Рисунок 2. Биологический цикл вегетации подсолнечника

Трансформация входных данных во входной датасет для модели осуществлялась по следующей схеме (см. рис. 3):

I. С помощью цикла, для каждого уникального ***meteoid*** и ***year*** создаем мини-датасет, где мин. индекс (i_start) начинается с недели на которой **t° почвы ≥8 °C**, а максимальный индекс (i_end) равен i_start+21.

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

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

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

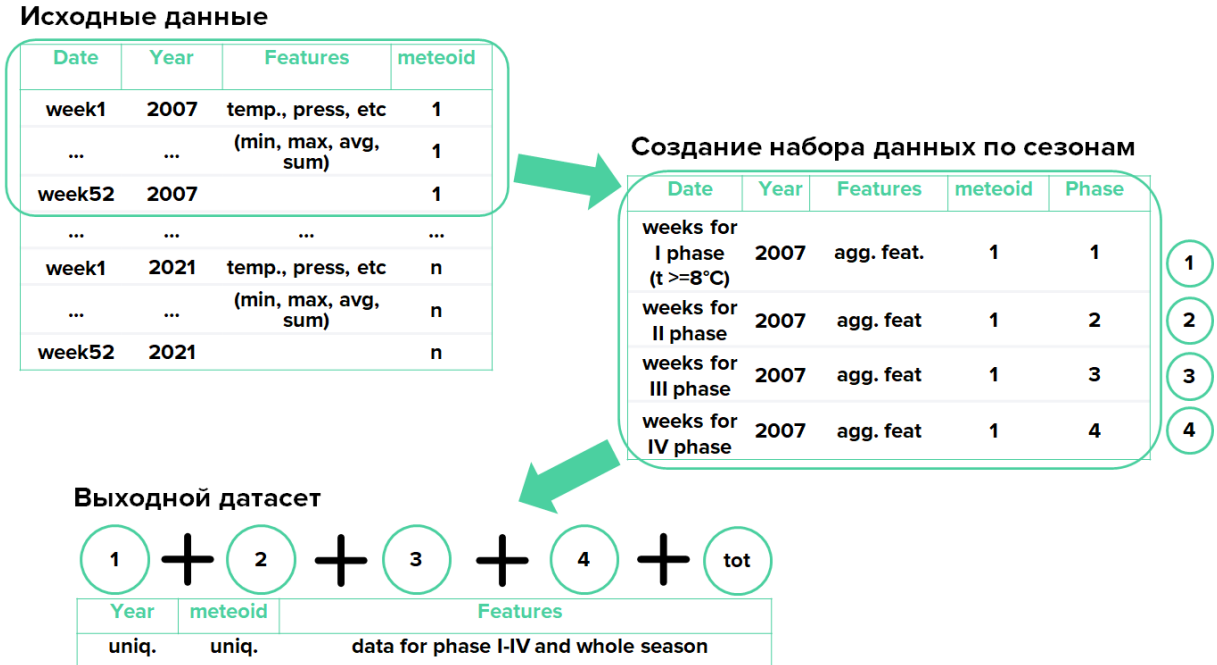
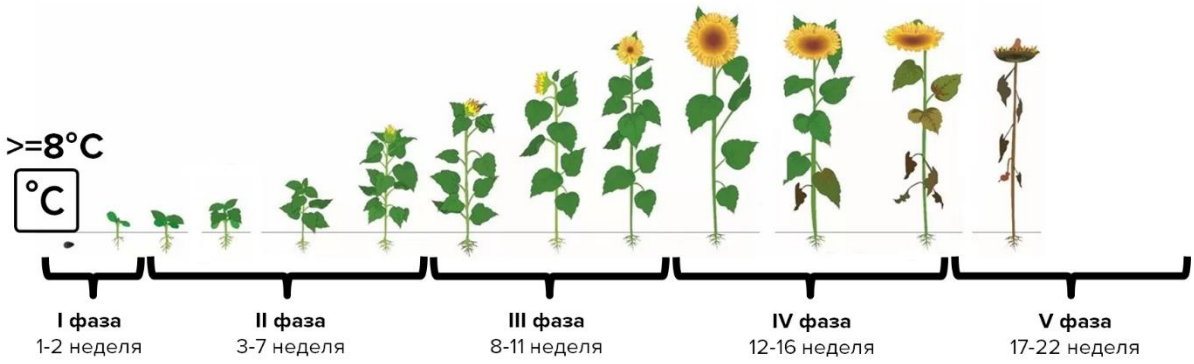


Рисунок 3. Схема трансформации исходных данных для создания входного вектора

Для обучения ***на вход в модель подается вектор признаков***, который содержит одинаковые переменные, сгруппированные по фазам вегетации, плюс признаки агрегированные за весь сезон (см. рис. 4). Целевой переменной выступает урожайность подсолнечника по муниципальному району за определенный год.



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	= target

Рисунок 4. Входной вектор признаков для обучения модели

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. Датасет с полным набором переменных «phase» содержит 165 переменных. (См. наборы данных «*total_data»* и «*weather*» из ноутбука "***preparation_mldata_4phase.ipynb***").

Во входной вектор признаков также были добавлены дополнительные features:

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

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

2. ***Суховей*** (ОПЯ[3] суховей) – feature name: dry_wind

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

3. ***Индикатор переувлажнения*** – feature name: precipitation_speed

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

4. ***Диапазон фичи*** (разница max-min) – features name: diff_<feature name>

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

5. ***Границы района*** – max, min и среднее значение геолокационных координат (долготы и широты) по муниципальному району, наименование features: <…>_border, central_lat, central_lon
6. ***Количество населенных пунктов*** в районе.
7. ***Севооборот*** – значение каждого 4-го года (от 0 - до 3)

[3] Опасное природное явление.

Суховей – ветер с высокой температурой, низкой относительной влажностью и большим дефицитом влажности летом в степях и полупустынях, направление – от восточного до южного. В агрометеорологической практике под суховеем обычно понимают ветер со скоростью более 5 м/с, при котором хотя бы в один срок наблюдений относительная влажность падает до 30% и ниже, температура воздуха повышается до 25° и выше.

Суховей вредно влияет на полевые культуры, так как при нем усиливается испарение и, при недостатке влаги в почве, верхние части растений увядают.

В процессе исследования были обучены несколько типов моделей: нейронные сети (LSTM Bidirectional, полносвязная нейронная сеть), модели градиентного бустинга (CatBoost Regressor, XGBoost Regressor), модель линейной регрессии (результаты обучения представлены в ноубуках: «***LSTM_learning_total.ipynb***», «***model_training_phase.ipynb***» и «***model_training_total.ipynb***»). Кроме того, для моделей бустинга также осуществлялся подбор оптимальных параметров (с помощью метода ***grid_search***). Итоговые результаты обучения моделей представлены в таблице 1.

Таблица 1 – Метрики качества по результатам обучения моделей на общих данных

Наименование модели	Значение RMSE test для наборов		Изм. RMSE test «phase» к «total»
	«total»	«phase»	
I. Нейронные сети			
LSTM Bidirectional	3.49	-	-
Полносвязная нейросеть	6.48	-	-
II. Регрессионные модели			
параметры default			
Linear Regression	114 676.876		
CatBoost Regression	2.811	2.747	-0.064
XGBoost Regression	2.147	2.119	-0.028
после подбора оптимальных параметров			
CatBoost Regression	2.635	2.164	-0.471
XGBoost Regression	2.022	1.979	-0.043

По результатам обучения моделей на общих данных можно сделать следующие выводы:

- Модель Linear Regression показала самые худшие результаты ($r^2 = 0.2$). В дальнейшем в исследованиях она не использовалась.
- Нейронные сети показывают более худшие результаты, чем бустинговые модели. Метрики качества нейронных сетей, полученные на тестовых выборках, показывают более низкие значения. Рекуррентная нейросеть показала более лучшие результаты, чем полносвязная нейросеть.
- Бустинговые модели показывают самые лучшие результаты. Метрики качества имеют более лучшие значения на датасете с полным набором признаков («phase»).
- Более лучшие результаты после обучения и оптимизации параметров на обоих датасетах показала модель XGB Regressor (RMSE test 1.979 для «phase»).

География регионов производства семян подсолнечника в России обширна. Кроме территориальных отличий есть также отличия по величине урожайности и объемам производства (см. табл. 2).

Таблица 2 – Производство семян подсолнечника по регионам России

Регион	**Средняя за 5 лет (2017-2021гг)**			**Средняя доля от, %**	
	Урожайность, ц/га	**Валовой сбор, тыс. т**	**Посевные площади, тыс. га**	**Валового сбора**	**Посевных площадей**
Российская Федерация	**16,2**	**12 973,1**	**8 607,2**	**100,0%**	**100,0%**
Приволжский ФО	12,9	4 671,9	3 837,4	36,0%	44,6%
Южный ФО	19,1	3 624,4	2 034,3	27,9%	23,6%
Центральный ФО	23,6	3 360,5	1 518,1	25,9%	17,6%
Сибирский ФО	10,2	665,0	745,9	5,1%	8,7%
Северо-Кавказский ФО	16,7	559,3	362,9	4,3%	4,2%
Уральский ФО	9,8	91,9	108,9	0,7%	1,3%
Дальневосточный ФО	9,8	0,2	0,5	0,0%	0,0%

Так, самое высокое значение средней урожайности (за последние 5 лет) отмечается в Центральном округе – 23,6 ц с га, самое низкое в Дальневосточном и Уральском округах – 9,8 ц с га. Самые большая доля производства приходится на Приволжский округ – 36,0%, самая низкая на Дальневосточный менее 0,01%. Таким образом, уровень урожайности по регионам имеет свои отличительные особенности и каждый регион имеет свой вес в общем объеме производства.

Чтобы учесть региональные особенности производства было произведено обучение моделей на наборах данных по каждому региону отдельно. Во-первых, это позволит моделям найти уникальные отличительные признаки по регионам для прогнозирования урожайности; во-вторых, позволит учитывать вес каждого региона при расчете итогового значения RMSE.

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

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

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

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

Было проведено обучение и тестирование моделей на наборах данных по 40 регионам. Совокупная доля производства подсолнечника в этих 40 регионах, за последние 5 лет, составила 99,8%. Таким образом, выборка, включающая указанные регионы, позволяет охватить практически все производство подсолнечника в России.

Результаты обучения моделей по регионам показывают, что взвешенное итоговое RMSE по регионам дает лучшее значение, чем простое среднее RMSE (см. табл. 3). Так, например, CatBoost reg. – значение $RMSE_{wg}$ = 1,498 (vs. RMSE = 2,273), XGBoost reg. – значение $RMSE_{wg}$ = 1,601 (vs. RMSE = 2,273). В случае если для отдельного региона использовать модель с лучшим значением RMSE, то возможно снизить итоговое значение $RMSE_{wg}$ до уровня 1,458.

Таблица 3 – Итоговые результаты обучения регрессионных моделей.

Параметр	**RMSE**	**RMSE к урожайности**	**$RMSE_{wg}$**	**$RMSE_{wg}$ к урожайности**
Средняя урожайность, ц/га	**16,2**	**-**	**-**	**-**
Лучшее значение из двух моделей	***2,273***[4]	***14,0%***	***1,458***	***9,0%***
CatBoost regressor	2,273	14,0%	1,498	9,3%
XGBoost regressor	2,273	14,0%	1,601	9,9%

Таким образом, за счет оптимизации способа расчета итогового значения метрики качества моделей удалось достигнуть целевого значения RMSE ≤ 10%. Даже при использовании моделей XGBoost, показывающих более худшие результаты для взвешенного RMSE, для прогнозирования урожайности итоговое значение $RMSE_{wg}$ не превысит 10%.

[4] Взято значение RMSE для модели XGBoost – 2,2729 (RMSE CatBoost = 2,2730)