Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико-механический институт

Работа допущена к защите
Руководитель ОП
_____ К.Н. Козлов
«_____» _____ 2025 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
РАБОТА БАКАЛАВРА
СРАВНЕНИЕ МОДЕЛЕЙ ЭФФЕКТИВНОСТИ ИНСЕКТИЦИДА

по направлению подготовки 01.03.02 Прикладная математика и информатика
Направленность (профиль) 01.03.02_02 Системное программирование

Выполнил
студент гр. 5030102/10201 А.П. Тасаков

Руководитель
доцент ВШПМиВФ,
к.ф.-м.н., с.н.с. С.В. Крашенинников

Консультант
доцент ВШПМиВФ, к.б.н. К.Н. Козлов

Консультант
по нормоконтролю Л.А. Арефьева

Санкт-Петербург — 2025

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт

УТВЕРЖДАЮ
Руководитель ОП
_____ К.Н. Козлов
«_____» _____ 2025г.

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

студенту Тасакову Антону Павловичу гр. 5030102/10201

1. Тема работы: Сравнение моделей эффективности инсектицида.
2. Срок сдачи студентом законченной работы: 28.05.2025.
3. Исходные данные по работе: основными источниками литературы являются статьи [3.1] и [3.2].
   3.1. Optimization of Flight Mode and Coupling Analysis of Operational Parameters on Droplet Deposition and Drift of Unmanned Aerial Spraying Systems (UASS) / Q. Liu [et al.] // Agronomy. — 2025. — Vol. 15. — P. 367. — DOI 10.3390/agronomy15020367.
   3.2. Optimizing Unmanned Aerial Vehicle Operational Parameters to Improve Pest Control Efficacy and Decrease Pesticide Dosage in Tea Gardens / M. Wu [et al.] // Agronomy. — 2025. — Vol. 15. — P. 431. — DOI 10.3390/agronomy15020431.
4. Содержание работы (перечень подлежащих разработке вопросов):
   4.1. Сбор и анализ данных из профильных научных статей.
   4.2. Исследование и отбор информативных признаков.
   4.3. Обучение и сравнение стандартных моделей машинного обучения.
   4.4. Обучение исследовательской модели и сравнение её с базовыми решениями.
5. Консультанты по работе:
   5.1. Доцент ВШПМиВФ, к.б.н., К.Н. Козлов.
   5.2. Ведущий программист, Л.А. Арефьева (нормоконтроль).
6. Дата выдачи задания: 31.01.2025.

Руководитель ВКР _____ С.В. Крашенинников

Консультант _____ К.Н. Козлов

Задание принял к исполнению 31.01.2025

Студент _____ А.П. Тасаков

РЕФЕРАТ

На 44 с., 16 рисунков, 2 таблицы, 0 приложений

КЛЮЧЕВЫЕ СЛОВА: РАСПЫЛЕНИЕ ИНСЕКТИЦИДОВ, БЕСПИЛОТНЫЕ ЛЕТАТЕЛЬНЫЕ АППАРАТЫ, МАШИННОЕ ОБУЧЕНИЕ, РЕГРЕССИОННЫЕ МОДЕЛИ, АГРОИНЖЕНЕРИЯ.

Тема выпускной квалификационной работы: «Сравнение моделей эффективности инсектицида».

ВКР посвящена применению методов машинного обучения для предсказания параметров распыления инсектицидов с использованием беспилотных летательных аппаратов (БПЛА). Цель исследования — построение и сравнение регрессионных моделей для прогнозирования площади покрытия и диаметра капель на основе агрегированных и аугментированных данных из открытых научных источников. В работе были использованы модели множественной линейной регрессии (MLR), регрессии опорных векторов (SVR), случайного леса (RF), градиентного бустинга (GBR), XGBoost и CatBoost. В результате CatBoost продемонстрировал наивысшую точность и стабильность. RF и GBR также показали хорошие, но уже менее устойчивые результаты. У XGBoost были отмечены признаки переобучения. MLR и SVR показали самые худшие результаты. Самыми важными признаками для обеих задач прогнозирования выступили объём распыления и скорость полёта. Также значительное влияние оказали режим атомизации, число форсунок и объём бака. Областью применения результатов являются агроинженерные системы, использующие БПЛА для обработки посевов. Модели машинного обучения показали высокую эффективность при наличии корректно подготовленных данных. Разработанные рекомендации могут быть использованы при проектировании систем поддержки агротехнологических решений.

ABSTRACT

44 pages, 16 figures, 2 tables, 0 appendices

KEYWORDS: INSECTICIDES SPRAYING, UNMANNED AERIAL VEHICLES, MACHINE LEARNING, REGRESSION MODELS, AGROENGINEERING.

The subject of the graduate qualification work is «Comparison of insecticide efficacy models».

This thesis is devoted to the application of machine learning methods for predicting insecticide spray parameters using unmanned aerial vehicles (UAVs). The aim of the study is to build and compare regression models for predicting coverage area and droplet diameter based on aggregated and augmented data from open scientific sources. The study used multiple linear regression (MLR), support vector regression (SVR), random forest (RF), gradient boosting (GBR), XGBoost, and CatBoost models. As a result, CatBoost demonstrated the highest accuracy and stability. RF and GBR also showed good, but less stable results. XGBoost showed signs of overfitting. MLR and SVR showed the worst results. The most important features for both prediction tasks were spray volume and flight speed. The atomization mode, number of nozzles, and tank volume also had a significant impact. The results can be applied to agricultural engineering systems that use UAVs for crop treatment. Machine learning models showed high efficiency when correctly prepared data was available. The recommendations developed can be used in the design of systems to support agricultural technology solutions.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Современное сельское хозяйство сталкивается с необходимостью повышения эффективности применения инсектицидов при одновременном снижении их негативного воздействия на растения и окружающую среду. В связи с этим особое внимание уделяется технологии обработки сельскохозяйственных культур с применением беспилотных летательных аппаратов (БПЛА) [1]. Возможность точной и гибкой настройки их параметров оказывает прямое влияние на эффективность доставки препарата к целевым поверхностям. При этом неправильно подобранные характеристики аппарата способны привести к избыточному расходу препарата, неравномерному покрытию, загрязнению окружающей среды и нанесению вреда обрабатываемым культурам [2]. Построение математических моделей, позволяющих прогнозировать данные параметры на основании входных условий, является важной прикладной задачей, решение которой может существенно повысить результативность агротехнических мероприятий.

На текущий момент существует множество исследований, посвящённых влиянию агротехнических и метеорологических условий на характеристики распыления при применении БПЛА. Например, в работе [3] рассмотрена задача подбора параметров работы БПЛА для уменьшения дрейфа и увеличения удержания раствора на листьях. Исследование [4] посвящено анализу условий применения инсектицидов в чайных плантациях. Аналогично была исследована эффективность распыления пестицида на полях с соей и пшеницей при использовании БПЛА с 4-мя типами форсунок [5], на кофейных плантациях с 3-мя разными сортами культуры [6] и в виноградниках [7]. Все эти исследования подчёркивают высокую зависимость результатов обработки от множества факторов: скорости и высоты полёта БПЛА, типа культуры, погодных условий и свойств раствора. Так, с биологической точки зрения актуальность работы заключается в необходимости повышения эффективности опрыскивания растений для борьбы с вредителями, болезнями и сорняками, которые сильно влияют на урожайность и качество продукции.

С другой стороны, все исследуемые процессы достаточно сложны для описания и поддаются формализации только с применением современных методов машинного обучения. На протяжении последних лет оно широко используется для анализа сложных, многомерных и зачастую зашумлённых данных. Его быстрое развитие позволяет применять современные алгоритмы анализа для построения предсказательных моделей во многих сферах нашей жизни. Например, в работах

[8] и [9] показано, что машинное обучение находит всё более широкое применение в сельском хозяйстве: от прогнозирования урожайности и определения состояния посевов до задач точного земледелия и оптимизации применения средств защиты растений. Тем не менее, для задачи прогнозирования параметров распыления инсектицида при помощи БПЛА сравнительный анализ эффективности различных моделей машинного обучения остаётся ограниченно представленным в научной литературе.

Объектом исследования является процесс опрыскивания сельскохозяйственных культур инсектицидами с помощью БПЛА. Предметом исследования выступают площадь покрытия (%) и диаметр капель (мкм), характеризующие эффективность распыления в зависимости от совокупности технических, агротехнических и метеорологических факторов.

Таким образом, цель работы: обучить и сравнить регрессионные модели, предназначенных для прогнозирования эффективности распыления инсектицида с применением БПЛА.

Для её достижения необходимо решить следующие задачи:

A. сформировать обучающий датасет на основе агрегированных и аугментированных данных из открытых научных публикаций;
B. обучить на подготовленном датасете следующие регрессионные модели: множественная линейная регрессия (MLR), регрессия опорных векторов (SVR), метод случайного леса (RF), градиентный бустинг (GBR), XGBoost и CatBoost;
C. провести оценку точности построенных моделей и сравнить их;
D. интерпретировать полученные результаты с позиции значимости входных факторов и применимости моделей в реальных условиях агроинженерии.

Для выполнения анализа в практической части ВКР были использованы материалы [3—7]. Из них были извлечены исходные данные для формирования датасета.

Методологической основой ВКР выступают методы машинного обучения, применяемые для построения и анализа регрессионных моделей. В частности: множественная линейная регрессия (MLR) [10], регрессия опорных векторов (SVR) [11; 12], случайный лес (RF) [13], градиентный бустинг (GBR) [14; 15], XGBoost [16] и CatBoost [17]. В работе используются методы предобработки и аугментации данных, а также подходы к оценке качества моделей на основе метрик регрессии.

ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ И МЕТОДИЧЕСКИЕ ОСНОВЫ ИССЛЕДОВАНИЯ

Глава посвящена изложению теоретических и прикладных аспектов, лежащих в основе построения моделей прогнозирования эффективности инсектицидной обработки с применением БПЛА.

В параграфе 1.1 формализуется задача работы. Параграф 1.2 описывает используемые методы машинного обучения. Рассматриваются как классические подходы (множественная линейная регрессия и регрессия опорных векторов), так и ансамблевые методы (случайный лес, градиентный бустинг, XGBoost и CatBoost), с кратким описанием их принципов работы. В параграфе 1.3 проводятся обзор и характеристика исходных данных. Описываются источники данных, их структура, ключевые признаки и целевые переменные, а также выполненные этапы предобработки данных для различных типов моделей. Параграф 1.4 описывает способы оценки моделей и алгоритмы настройки их гиперпараметров.

1.1. Постановка задачи

Рассмотрим задачу построения прогностических моделей, направленных на оценку эффективности опрыскивания сельскохозяйственных культур инсектицидами при помощи БПЛА. В рамках данного исследования предполагается, что на распыление существенное влияние оказывают следующие группы факторов:

- характеристики БПЛА, включая параметры его конструкции, высоту полёта, скорость и тип форсунок;
- свойства целевой сельскохозяйственной культуры (тип культуры, морфология листьев и пр.);
- погодные условия во время обработки (скорость ветра, температура воздуха и влажность).

Целевыми переменными в рамках данной задачи выступают:

- $y_1 \in [0, 100] \subset \mathbb{R}$ — процент площади покрытия растительности рабочей жидкостью (далее — *покрытие*);
- $y_2 \in \mathbb{R}_+$ — средний диаметр капель распыления, выраженный в микрометрах (далее — *размер капель*).

Пусть имеется выборка наблюдений, составленная на основе агрегированных и аугментированных данных, извлечённых из открытых научных публикаций

[3; 4], в которых представлены экспериментальные измерения эффективности опрыскивания при различных конфигурациях БПЛА, типах культур и внешних условиях. Формально, обозначим выборку $\mathcal{D}$ размера N как:

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, y_1^{(i)}, y_2^{(i)} \right) \right\}_{i=1}^{N}, \tag{1.1}$$

где каждый $\mathbf{x}^{(i)} \in \mathbb{R}^d$ — вектор признаков, описывающий конкретную реализацию комбинации параметров:

$$\mathbf{x}^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, \ldots, x_d^{(i)} \right], \tag{1.2}$$

включающий как числовые переменные (скорость полёта, температура воздуха и т.п.), так и категориальные (тип форсунки, сорт культуры и т.п.). Размерность пространства признаков d определяется количеством отобранных характеристик. Для упрощения формулировки будем считать, что все категориальные признаки закодированы в виде one-hot или target encoding и подлежат обработке с помощью моделей машинного обучения.

Необходимо построить две аппроксимирующие модели $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$, такие что:

$$\hat{y}_1 = f_1(\mathbf{x}), \quad \hat{y}_2 = f_2(\mathbf{x}), \tag{1.3}$$

где $\hat{y}_1$ и $\hat{y}_2$ — предсказанные моделью процент покрытия и средний диаметр капель соответственно. Требуется, чтобы модели минимизировали отклонение между предсказанными и фактическими значениями на тестовой выборке с учётом выбранной функции потерь.

Обучающая выборка сформирована путём объединения результатов независимых исследований, найденных в научной литературе [3; 4; 6], с последующим применением процедур синтетического расширения (data augmentation), включая стохастическое варьирование условий и моделирование шумов измерения.

Таким образом, задача направлена на решение многомерной регрессии с двумя зависимыми переменными, на входе которой — вектор признаков, характеризующих совокупность агротехнических, технических и метеорологических условий, а на выходе — количественные оценки параметров эффективности опрыскивания, подлежащие интерпретации в прикладных агроинженерных задачах.

В ходе работы применялись и сравнивались несколько классов моделей: метод множественной линейной регрессии (MLR), регрессия опорных векторов

(SVR), ансамблевые методы — случайный лес (RF), градиентный бустинг (GBR), XGBoost и CatBoost.

1.2. Обзор используемых методов

1.2.1. Множественная линейная регрессия (MLR)

Множественная линейная регрессия (multiple linear regression, MLR) — один из базовых методов машинного обучения и статистического анализа, предназначенный для моделирования зависимости одной количественной переменной от нескольких независимых признаков. В современном виде теория MLR была подробно изложена в работах Роналда Фишера и Джона Неймена [10; 18].

1.2.1.1. Формализация задачи

Предполагается, что между признаками и откликом существует линейная зависимость следующего вида:

$$y^{(i)} = \beta_0 + \sum_{j=1}^{d} \beta_j x_j^{(i)} + \varepsilon^{(i)}, \quad i = 1, \ldots, N, \tag{1.4}$$

где $\beta_0 \in \mathbb{R}$ — свободный член (смещение), $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_d]^\top \in \mathbb{R}^d$ — вектор коэффициентов линейной модели, $\varepsilon^{(i)}$ — ошибка (шум), предполагаемая независимой и нормально распределённой случайной величиной с нулевым математическим ожиданием и постоянной дисперсией: $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$.

Цель метода — найти такие параметры β_0, $\boldsymbol{\beta}$, которые минимизируют сумму квадратов отклонений модели от наблюдаемых значений, т.е. решить задачу минимизации функции ошибки:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{N} \left(y^{(i)} - \beta_0 - \sum_{j=1}^{d} \beta_j x_j^{(i)} \right)^2. \tag{1.5}$$

Обозначим через $\mathbf{X} \in \mathbb{R}^{N \times (d+1)}$ матрицу признаков с добавленным столбцом единиц (для учёта β_0), и через $\mathbf{y} \in \mathbb{R}^N$ — вектор откликов. Тогда модель можно записать в матричной форме:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.6}$$

где $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_d]^\top \in \mathbb{R}^{d+1}$, и задача минимизации принимает форму:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2. \tag{1.7}$$

Решение данной задачи (при условии, что $\mathbf{X}^\top\mathbf{X}$ невырождено) задаётся аналитически формулой нормальных уравнений:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{1.8}$$

1.2.1.2. Преимущества и ограничения

Полученная модель обладает рядом свойств: несмещённость оценки, минимальная дисперсия среди всех линейных несмещённых оценок (согласно теореме Гаусса—Маркова) и возможность проведения статистических тестов на значимость коэффициентов [10; 18].

MLR активно применяется в задачах прогноза, интерпретируемого анализа данных и оценки влияния факторов, где линейная зависимость может считаться допустимым приближением. При этом модель чувствительна к мультиколлинеарности и выбросам, что требует предварительной проверки условий применимости, таких как полнота ранга матрицы признаков и гомоскедастичность остатков.

1.2.2. Регрессия опорных векторов (SVR)

Регрессия опорных векторов (Support Vector Regression, SVR) представляет собой обобщение метода опорных векторов (Support Vector Machines, SVM), изначально разработанного для задач классификации, на случай задач регрессии. Теоретическая основа метода была заложена в работах Вапника и др. [11; 12].

В отличие от классических методов, ориентированных на минимизацию эмпирической ошибки, метод опорных векторов стремится к оптимальному компромиссу между точностью на обучающей выборке и сложностью модели, что позволяет достичь высокой обобщающей способности, особенно в условиях ограниченных данных или высокой размерности пространства признаков.

1.2.2.1. Формализация задачи

В SVR предполагается, что модель f имеет вид:

$$f(\mathbf{x}) = \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle + b, \tag{1.9}$$

где $\varphi : \mathbb{R}^d \to \mathcal{H}$ — отображение исходного пространства признаков в высокоразмерное (возможно бесконечномерное) гильбертово пространство признаков, $\mathbf{w} \in \mathcal{H}$ — вектор весов, $b \in \mathbb{R}$ — смещение, а $\langle \cdot, \cdot \rangle$ — скалярное произведение в $\mathcal{H}$.

Вводится ε-интенсивная зона нечувствительности, в пределах которой отклонение предсказания от истинного значения не штрафуется. Тогда задача SVR формулируется как задача минимизации функционала:

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*), \tag{1.10}$$

при ограничениях:

$$\begin{cases} y^{(i)} - \langle \mathbf{w}, \varphi(\mathbf{x}^{(i)}) \rangle - b \leqslant \varepsilon + \xi_i, \\ \langle \mathbf{w}, \varphi(\mathbf{x}^{(i)}) \rangle + b - y^{(i)} \leqslant \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geqslant 0, \quad i = 1, \ldots, N, \end{cases} \tag{1.11}$$

где ξ_i, ξ_i^* — вспомогательные переменные, отвечающие за отклонения вне ε-зоны, а $C > 0$ — гиперпараметр, контролирующий компромисс между допустимой величиной ошибок и гладкостью модели.

Переходя к двойственной задаче, получаем квадратичную задачу оптимизации:

$$\max_{\alpha, \alpha^*} -\frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) y^{(i)} - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i^*), \tag{1.12}$$

при ограничениях:

$$\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0, \quad 0 \leqslant \alpha_i, \alpha_i^* \leqslant C. \tag{1.13}$$

Здесь $K(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$ — положительно определённая ядерная функция (линейное, полиномиальное, сигмоидальное или радиальное базисное (Гауссовское) ядро). Решение задачи позволяет выразить модель в виде:

$$f(\mathbf{x}) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(\mathbf{x}^{(i)}, \mathbf{x}) + b. \tag{1.14}$$

Ненулевые множители $(\alpha_i - \alpha_i^*)$ соответствуют *опорным векторам* — наблюдениям, лежащим за пределами ε-зоны или на её границах. Это придаёт методу свойство разреженности и высокую эффективность в больших размерностях.

1.2.2.2. Преимущества и ограничения

Метод SVR обладает высокой устойчивостью к переобучению, особенно в условиях малых выборок или высокой коррелированности признаков, благодаря геометрической интерпретации и контролю сложности модели через норму весов $\|\mathbf{w}\|$. Однако он чувствителен к выбору параметров C, ε, а также типа и параметров ядра. Решение двойственной задачи требует квадратичного программирования, что может быть вычислительно дорого при больших объёмах данных.

1.2.3. Метод случайного леса (RF)

Метод случайного леса (Random Forest, RF) представляет собой ансамблевый алгоритм машинного обучения, предложенный Лео Брейманом в 2001 году [13]. Он объединяет концепции бэггинга (bootstrap aggregating) и случайного выбора подмножества признаков при построении каждого узла дерева, что позволяет создавать устойчивые и высокообобщающие модели для задач классификации и регрессии.

1.2.3.1. Формализация метода

Метод случайного леса строит ансамбль из B независимых регрессионных деревьев $\{T_b(\mathbf{x})\}_{b=1}^{B}$, каждое из которых обучается на бутстрэп-подвыборке $\mathcal{D}_b \subseteq \mathcal{D}$, полученной путём случайного выбора объектов с возвращением. Затем строится решение дерева T_b по следующему алгоритму: в каждом узле вместо полного множества признаков $\{1, \ldots, d\}$ рассматривается случайное подмножество индексов $\mathcal{M}_b \subset \{1, \ldots, d\}$ размера $m \ll d$, и сплит выбирается по признаку из $\mathcal{M}_b$, минимизирующему критерий разбиения. Итоговое предсказание в задаче регрессии задаётся усреднением по ансамблю:

$$f(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}). \tag{1.15}$$

Каждое дерево T_b строится до максимальной глубины без обрезки, делая его низкосмещённым, но сильно переобученным. Усреднение по ансамблю снижает дисперсию.

Случайный характер отбора наблюдений и признаков приводит к снижению корреляции между деревьями, что является критическим фактором в обеспечении хороших обобщающих свойств ансамбля. Как показал Брейман [13], уменьшение корреляции между базовыми моделями при сохранении их точности снижает общее отклонение ансамбля:

$$\mathrm{Var}(f(\mathbf{x})) = \rho \cdot \sigma^2 + \frac{1-\rho}{B} \cdot \sigma^2, \tag{1.16}$$

где ρ — средняя корреляция между базовыми деревьями, σ^2 — их дисперсия. При $B \to \infty$, второй член стремится к нулю, и обобщающая ошибка определяется корреляцией и индивидуальной дисперсией деревьев.

1.2.3.2. Преимущества и ограничения

Метод случайного леса обладает высокой устойчивостью к переобучению, не требует масштабирования признаков и хорошо справляется как с линейно, так и с нелинейно разделимыми зависимостями. Он эффективно работает с выборками высокой размерности и может использоваться для оценки важности признаков.

Однако модель теряет интерпретируемость по сравнению с одиночными деревьями и может быть ресурсоёмкой при большом числе деревьев и глубине.

1.2.4. Градиентный бустинг (GBR)

Градиентный бустинг (Gradient Boosting Regression, GBR) — это ансамблевый метод построения прогнозирующих моделей, основанный на последовательной комбинации слабых предсказателей, как правило, решающих деревьев, с целью повышения точности регрессии. Метод был впервые формализован Джеромом Фридманом в начале 2000-х годов как обобщение алгоритма AdaBoost на произвольные дифференцируемые функции потерь [14].

Градиентный бустинг реализует стратегию жадного приближения, при которой каждая последующая модель обучается на ошибках предыдущей, минимизируя выбранную функцию потерь с использованием методов градиентного спуска в функциональном пространстве.

1.2.4.1. Формализация задачи

Цель метода — найти функцию $F(\mathbf{x})$, аппроксимирующую зависимость между признаками и откликом путём минимизации эмпирического риска:

$$F^*(\mathbf{x}) = \arg\min_{F \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(y^{(i)}, F(\mathbf{x}^{(i)})), \tag{1.17}$$

где $\mathcal{L}$ — дифференцируемая по второму аргументу функция потерь, а $\mathcal{F}$ — класс допустимых функций (например, пространство деревьев решений).

Идея градиентного бустинга заключается в поэтапном приближении решения через функциональный градиентный спуск. На этапе m при известной текущей оценке $F_{m-1}(\mathbf{x})$ вычисляются «псевдо-остатки» — отрицательные градиенты по предсказаниям:

$$r_i^{(m)} = - \left. \frac{\partial L\left(y^{(i)}, F(\mathbf{x}^{(i)})\right)}{\partial F(\mathbf{x}^{(i)})} \right|_{F=F_{m-1}}. \tag{1.18}$$

Затем обучается базовый регрессор $h_m(\mathbf{x})$ на данных $\{(\mathbf{x}^{(i)}, r_i^{(m)})\}$, то есть решается:

$$h_m = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{N} \left(r_i^{(m)} - h(\mathbf{x}^{(i)})\right)^2. \tag{1.19}$$

Далее находится оптимальный шаг по линии:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{N} L\left(y^{(i)}, F_{m-1}(\mathbf{x}^{(i)}) + \gamma\, h_m(\mathbf{x}^{(i)})\right), \tag{1.20}$$

и модель обновляется по правилу:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu\, \gamma_m\, h_m(\mathbf{x}), \tag{1.21}$$

где $\nu \in (0,1]$ — параметр скорости обучения (*learning rate*), снижающий размер шага для повышения устойчивости.

1.2.4.2. Преимущества и ограничения

Градиентный бустинг сочетает в себе сильные стороны ансамблевых методов и градиентного спуска, обеспечивая высокую предсказательную способность даже при сложных нелинейных зависимостях и смешанных типах признаков.

Благодаря параметру ν и ограничению глубины деревьев можно контролировать сложность модели и снижать риск переобучения, а последовательная жадная схема обеспечивает интерпретируемость вклада каждого добавленного дерева [15]. Вместе с тем метод чувствителен к настройке гиперпараметров: число итераций, скорость обучения, максимальная глубина базовых деревьев и критерий остановки должны подбираться тщательно, зачастую с помощью перекрёстной проверки. Кроме того, из-за итеративного обучения ансамбля из M компонентов вычислительная и пространственная сложности метода растут линейно с M, что может стать ограничением при больших объёмах данных или при необходимости быстрой онлайн-обучаемости [15].

1.2.5. XGBoost

Метод XGBoost (Extreme Gradient Boosting) — одна из наиболее производительных реализаций градиентного бустинга (см. п. 1.2.4), разработанную Тяньци Ченом и Карлосом Гуэстрином [16]. Он был предложен как масштабируемое и регуляризованное обобщение градиентного бустинга, способное эффективно работать с большими разреженными наборами данных и обеспечивать высокое качество модели при относительно низких затратах вычислительных ресурсов.

1.2.5.1. Формализация метода

Модель XGBoost строится в виде суммы K аддитивных базовых моделей $f_k \in \mathcal{F}$, где $\mathcal{F}$ — пространство деревьев решений:

$$\hat{y}^{(i)} = \sum_{k=1}^{K} f_k(\mathbf{x}^{(i)}), \quad f_k \in \mathcal{F}. \tag{1.22}$$

Каждое дерево f_k представляет собой структуру, сопоставляющую входному признаковому вектору значение на одном из листьев дерева. Обучение осуществляется путём последовательного добавления новых деревьев, уменьшающих значение заданной функции потерь.

Функция оптимизации имеет вид:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l\left(y^{(i)}, \hat{y}^{(i)(t-1)} + f_t(\mathbf{x}^{(i)})\right) + \Omega(f_t), \tag{1.23}$$

где l — дифференцируемая функция потерь (например, среднеквадратичная ошибка), $\Omega(f)$ — регуляризационный член, накладывающий штраф за сложность модели:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{1.24}$$

где T — число листьев в дереве, w_j — вес j-го листа, γ и λ — гиперпараметры регуляризации. Такой подход позволяет контролировать переобучение за счёт структурной регуляризации.

Для эффективной оптимизации функции потерь применяется второй порядок разложения Тейлора:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N} \left[l(y^{(i)}, \hat{y}^{(i)(t-1)}) + g_i f_t(\mathbf{x}^{(i)}) + \frac{1}{2} h_i f_t^2(\mathbf{x}^{(i)}) \right] + \Omega(f_t), \tag{1.25}$$

где $g_i = \partial_{\hat{y}^{(i)(t-1)}} l(y^{(i)}, \hat{y}^{(i)(t-1)})$ и $h_i = \partial^2_{\hat{y}^{(i)(t-1)}} l(y^{(i)}, \hat{y}^{(i)(t-1)})$ — первая и вторая производные функции потерь.

Суммарный прирост качества от добавления нового дерева рассчитывается по формуле:

$$\mathcal{L}_{\text{split}} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma, \tag{1.26}$$

где G_L, H_L и G_R, H_R — суммы градиентов и гессианов для левого и правого поддеревьев соответственно. Выбор лучшего разделения узла осуществляется по максимальному приросту функции.

1.2.5.2. Преимущества и ограничения

Ключевой особенностью XGBoost является наличие регуляризации, которая способствует снижению вероятности переобучения и улучшает обобщающую способность модели. Благодаря блочной структуре представления данных, параллельных вычислений при построении деревьев и буферизации промежуточных градиентов и гессианов, алгоритм демонстрирует высокую вычислительную эффективность, особенно при работе с большими объёмами данных.

Однако, метод чувствителен к выбору гиперпараметров, от которых существенно зависит точность и устойчивость модели. При избыточной глубине

деревьев или слишком большом числе итераций возможно переобучение, особенно при наличии шумов в данных. Кроме того, несмотря на наличие различных оптимизаций, при работе с действительно большими и разреженными наборами данных объём вычислений и требования к памяти могут быть значительными, что накладывает ограничения на применимость в условиях ограниченных ресурсов.

1.2.6. CatBoost

CatBoost (Categorical Boosting) — это модификация градиентного бустинга на решающих деревьях, специально ориентированная на эффективную обработку категориальных признаков и устойчивость к переобучению. Метод разработан исследовательской группой компании Яндекс под руководством Прохоренко Л. и Гусева А. [17]. CatBoost отличается от других реализаций бустинга использованием упорядоченного бустинга, схематически предотвращающего утечку информации, и особой техникой кодирования категориальных переменных на основе таргет-статистик с контролем смещения.

1.2.6.1. Формализация метода

CatBoost вводит ключевую модификацию стандартного бустинга — Ordered Boosting, направленную на устранение предвзятой оценки градиента, возникающей при использовании текущих предсказаний модели для оценки градиентов на тех же данных.

Вместо использования всей обучающей выборки для расчёта градиентов, CatBoost разбивает данные на набор перестановок $\sigma \in \mathfrak{S}_n$ и для каждого объекта x_i использует лишь предсказания, полученные на подвыборке из объектов, предшествующих ему в перестановке σ:

$$\hat{g}_i^{(m)} = \left. \frac{\partial L(y_i, \hat{F}_{m-1}(x_i))}{\partial \hat{F}} \right|_{\hat{F}=\sum_{j\in\{j:\sigma(j)<\sigma(i)\}} f_j(x_j)} \tag{1.27}$$

CatBoost использует оригинальный метод преобразования категориальных признаков, основанный на таргет-статистике с порядком. В отличие от обычного one-hot-encoding, CatBoost рассчитывает условное ожидание целевой переменной по значению категориального признака с использованием техники «leave-one-out» в упорядоченной последовательности:

$$\text{Stat}(x_i^j) = \frac{\sum_{k:\sigma(k)<\sigma(i)} \mathbf{1}[x_k^j = x_i^j] \cdot y_k + a \cdot \mu}{\sum_{k:\sigma(k)<\sigma(i)} \mathbf{1}[x_k^j = x_i^j] + a} \quad (1.28)$$

где a — параметр сглаживания, μ — глобальное среднее значения целевой переменной.

1.2.6.2. Преимущества и ограничения

Преимущества CatBoost заключаются в его способности эффективно обрабатывать категориальные признаки без необходимости ручного кодирования. Алгоритм устраняет смещение в оценке градиентов, что снижает переобучение и улучшает обобщающую способность модели, особенно на небольших и разреженных выборках.

Ограничения CatBoost связаны в первую очередь с его вычислительной сложностью: по сравнению с другими реализациями бустинга он может требовать больше времени на обучение, особенно на очень больших наборах данных.

1.3. Обзор и характеристика исходных данных

В рамках исследования использовались данные, собранные из трёх научных публикаций, каждая из которых описывает экспериментальные применения БПЛА для инсектицидной обработки сельскохозяйственных культур. Первая статья [4] посвящена исследованию влияния параметров полёта (высоты, скорости, объёма распыления и режима автоматизации) на эффективность обработки в чайных плантациях с использованием двух моделей дронов: AGRAS DT30 и T40. Вторая [3] — анализирует взаимосвязь между параметрами работы БПЛА и такими показателями, как дрейф капель и эффективность осаждения рабочей жидкости на полях с пшеницей. Третья [6] — рассматривает влияние высоты полёта и морфологических особенностей различных генотипов кофейных растений на параметры распыления.

Первая статья содержит 24 наблюдения, включающих 30 признаков и две целевые переменные. Вторая и третья — предоставляют по 9 наблюдений каждая с 24 и 25 признаками соответственно, также с двумя целевыми переменными. Таким образом, совокупный массив составил 42 наблюдения, охватывающих разнообразные характеристики плантаций, параметры используемой техники, условия проведения эксперимента и результаты обработки.

Исходные признаки охватывают 4 смысловых блока. Первый блок описывает характеристики обрабатываемых культур, включая сорт, название и генотип растения, а также его особенности — среднюю высоту, диаметр полога, площадь листьев и другие морфологические показатели. Второй блок посвящён техническим параметрам применяемых БПЛА. В частности, учитываются модель аппарата, масса, объём бака, габаритные размеры (длина, ширина, высота), характеристики пропеллеров (мощность и количество), параметры форсунок (модель, количество, диапазон диаметров капель и ширины захвата), а также конфигурация водяного насоса (модель, максимальный расход, число насосов). Третий блок описывает условия эксперимента: погодные параметры (температура, влажность воздуха, скорость ветра) и настраиваемые параметры полёта — рабочая высота, скорость и объём опрыскивания на единицу площади. Наконец, в четвёртом блоке представлены целевые переменные, по которым оценивалась эффективность обработки: площадь покрытия рабочей жидкостью (%) и средний размер образующихся капель (мкм).

С целью последующего анализа и обучения моделей были сформированы три версии датасета. Первая — незакодированный массив с сохранением категориальных признаков и нулевых значений — предназначалась для использования в модели CatBoost, обладающей встроенной поддержкой таких данных. Размерность этой версии составила 420×40 для задачи прогнозирования площади покрытия и 360×40 — для задачи оценки размера капель. Вторая — закодированный датасет с предварительным удалением неинформативных признаков (нулевых столбцов) и использовалась при обучении моделей XGBoost, Gradient Boosting Regressor и случайного леса. Для обеих задач она имела размерность 420×28 и 360×28 соответственно. Третья — дополнительно масштабированный закодированный датасет — применялась для множественной линейной регрессии и регрессии опорных векторов. Её структура совпадала со второй версией.

1.4. Методы оценки качества моделей и стратегии настройки гиперпараметров

1.4.1. Метрики качества регрессионных моделей

Оценка эффективности моделей машинного обучения требует использования количественных метрик, отражающих степень соответствия предсказаний

реальным значениям. В рамках данного исследования применялись две метрики: коэффициент детерминации (R^2) и среднеквадратичная ошибка ($RMSE$).

Коэффициент детерминации R^2 измеряет долю дисперсии зависимой переменной, объяснённую моделью. Он определяется как:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1.29}$$

где y_i — наблюдаемые значения, $\hat{y}_i$ — предсказанные моделью значения, а $\bar{y}$ — среднее значение зависимой переменной. $R^2 \in [0, 1]$ интерпретируется как доля вариации, объяснённая моделью. При $R^2 \to 1$ предсказания с большой точностью аппроксимируют истинные значения. Отрицательные значения возможны при неадекватности модели.

Среднеквадратичная ошибка (Root Mean Squared Error, $RMSE$) измеряет абсолютное среднее отклонение предсказаний от истинных значений:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{1.30}$$

Эта метрика имеет ту же размерность, что и зависимая переменная, что делает её интерпретируемой в прикладных задачах. $RMSE$ чувствителен к выбросам.

1.4.2. Метод k-кратной кросс-валидации

Для получения устойчивой и менее смещённой оценки обобщающей способности модели применялась k-кратная кросс-валидация (k-fold cross-validation). В этой процедуре исходная обучающая выборка разбивается на k равных (или почти равных) блоков. На каждой итерации один из них используется в качестве валидационного множества, а оставшиеся $k - 1$ — для обучения модели. После завершения всех k итераций рассчитываются значения метрик качества, усреднённые по всем блокам. Такой подход позволяет выявить переобучение и оценить устойчивость модели к варьированию входных данных [15]. Часто выбираются значения $k = 5$ или $k = 10$, обеспечивающие баланс между вычислительной нагрузкой и статистической надёжностью.

1.4.3. Методы подбора гиперпараметров

1.4.3.1. Подбор параметров методом перебора по сетке

Grid Search представляет собой исчерпывающий перебор всех возможных комбинаций значений гиперпараметров, заданных в виде дискретных сеток. Для каждой комбинации обучается модель, после чего её качество оценивается с использованием кросс-валидации. Итоговая конфигурация выбирается как та, что демонстрирует наилучшее значение метрики (максимальное R^2 или минимальное $RMSE$). Метод является детерминированным и гарантирует нахождение глобального оптимума в пределах заданного пространства поиска, однако при увеличении размерности пространства экспоненциально увеличивается и сложность поиска.

1.4.3.2. Подбор параметров с помощью генетического алгоритма

Генетический алгоритм — это стохастическая эвристическая стратегия оптимизации, вдохновлённая принципами естественного отбора и эволюции. Каждый набор гиперпараметров кодируется как «индивид» (или «хромосома»), а множество таких конфигураций образует «популяцию». Алгоритм начинается с инициализации случайной популяции. На каждом поколении происходит:

A. *Оценка приспособленности (fitness evaluation)*: вычисление метрики качества (например, RMSE на валидации) для каждого индивида.
B. *Селекция*: отбор наиболее приспособленных особей, вероятность выбора которых пропорциональна их приспособленности.
C. *Кроссовер (скрещивание)*: комбинирование гиперпараметров двух родительских индивидов для генерации потомков.
D. *Мутация*: случайная модификация отдельных параметров потомков с небольшой вероятностью, обеспечивающая исследование пространства.
E. *Замещение*: формирование новой популяции, состоящей из лучших особей и новых потомков.

Процесс повторяется на протяжении фиксированного числа поколений или до достижения сходимости. Генетический алгоритм способен находить хорошие решения в высокоразмерных и сложных пространствах поиска, где другие методы теряют эффективность [19].

1.5. Выводы

Формальная постановка задачи позволила определить целевые переменные и структуру пространства признаков. Обзор методов машинного обучения продемонстрировал преимущества и ограничения выбранных алгоритмов для решения задачи регрессии. Характеристика данных описала структуру признаков и три версии датасета, адаптированные под особенности каждой группы моделей. С помощью метрик и методов оценки моделей была обеспечена объективность сравнения и стабильность настройки моделей, чувствительных к гиперпараметрам. Таким образом, сформирована методическая база для последующего сравнительного анализа эффективности выбранных алгоритмов.

ГЛАВА 2. ЭТАПЫ ПОДГОТОВКИ И АНАЛИЗА ДАННЫХ

В этой главе описываются этапы расширения и унификации экспериментальных данных, необходимых для обучения моделей. В параграфе 2.1 исходные таблицы аугментируются с сохранением исходных статистических свойств. Затем в параграфе 2.2 они объединяются в единый набор, и после анализа пропусков формируются три итоговые версии датасета: для CatBoost, для других ансамблевых методов (XGBoost, RF, GBR) и для MLR и SVR.

2.1. Аугментация экспериментальных данных

Для повышения обобщающей способности прогнозных моделей и расширения пространства признаков, доступных на этапе обучения, была проведена аугментация исходных данных. Основной целью процедуры являлось создание дополнительных наблюдений, сохраняющих ключевые статистические характеристики целевых переменных, но варьирующих параметры в пределах допустимого диапазона. Такой подход особенно оправдан при ограниченном объёме эмпирических данных, полученных из публикаций, где количество наблюдений зачастую невелико, а разнообразие условий проведения экспериментов недостаточно для обучения устойчивых моделей машинного обучения [20; 21].

Аугментация выполнялась отдельно для каждого датасета, после чего формировались два дополнительных набора: с увеличением объёма данных в пять и в десять раз по сравнению с оригинальным. Варьирование признаков осуществлялось при помощи стохастического отклонения значений отдельных параметров (например, морфологических характеристик растения, метеоусловий, высоты полёта и скорости обработки) в пределах заранее заданных границ. Они определялись либо на основе допусков, зафиксированных в первоисточниках [3; 4; 6], либо на предварительном анализе разброса по исходным данным. Для различных признаков варьирование осуществлялось с разной интенсивностью, например, параметры метеоусловий и опрыскивания допускали большую степень отклонения, чем морфологические признаки растений.

Для валидации корректности применённой процедуры аугментации были построены сравнительные графики целевых метрик — покрытия поверхности растения рабочей жидкостью (coverage, в %) и среднего диаметра капель (droplet size, в мкм). На рис.2.1 и рис.2.2 представлены результаты для оригинальных и

аугментированных выборок (с коэффициентами увеличения ×5 и ×10), рассчитанные отдельно для каждого из экспериментальных источников: AGRAS T30 и T40 по данным [4], AGRAS T20 по [3], JT5L-404 по [6].

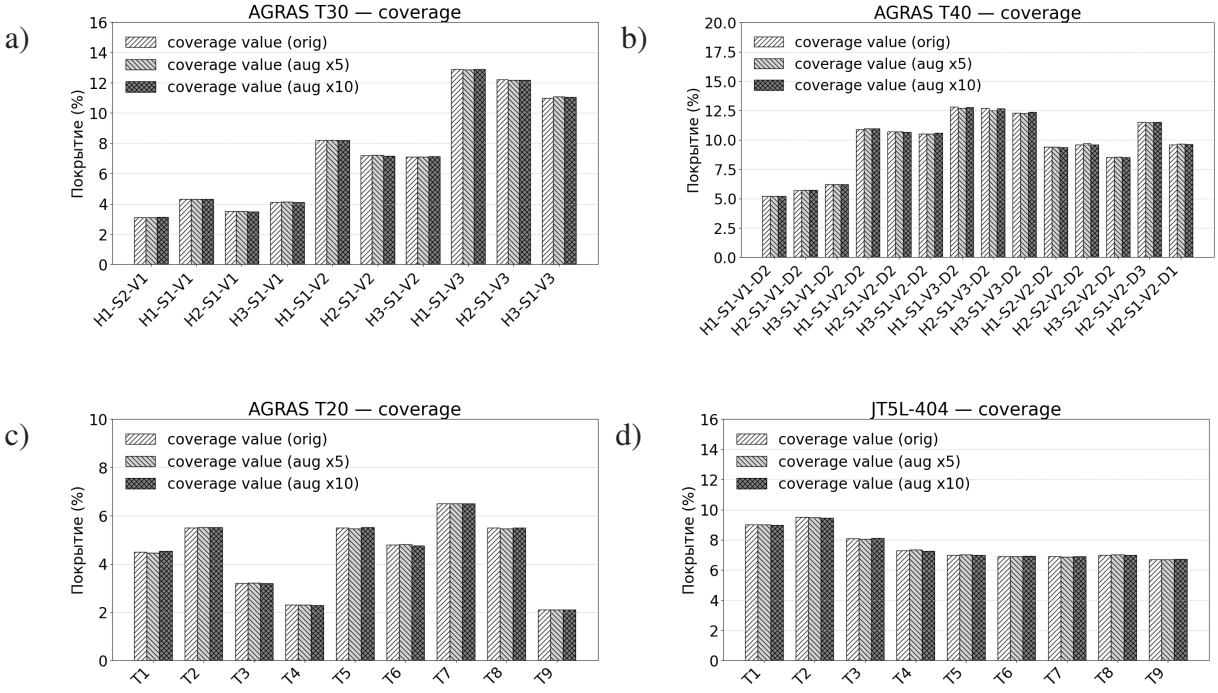

Рис.2.1. Распределение значений процента покрытия в исходной и аугментированных выборках: *a* — AGRAS T30; *b* — AGRAS T40; *c* — AGRAS T20; *d* — JT5L-404

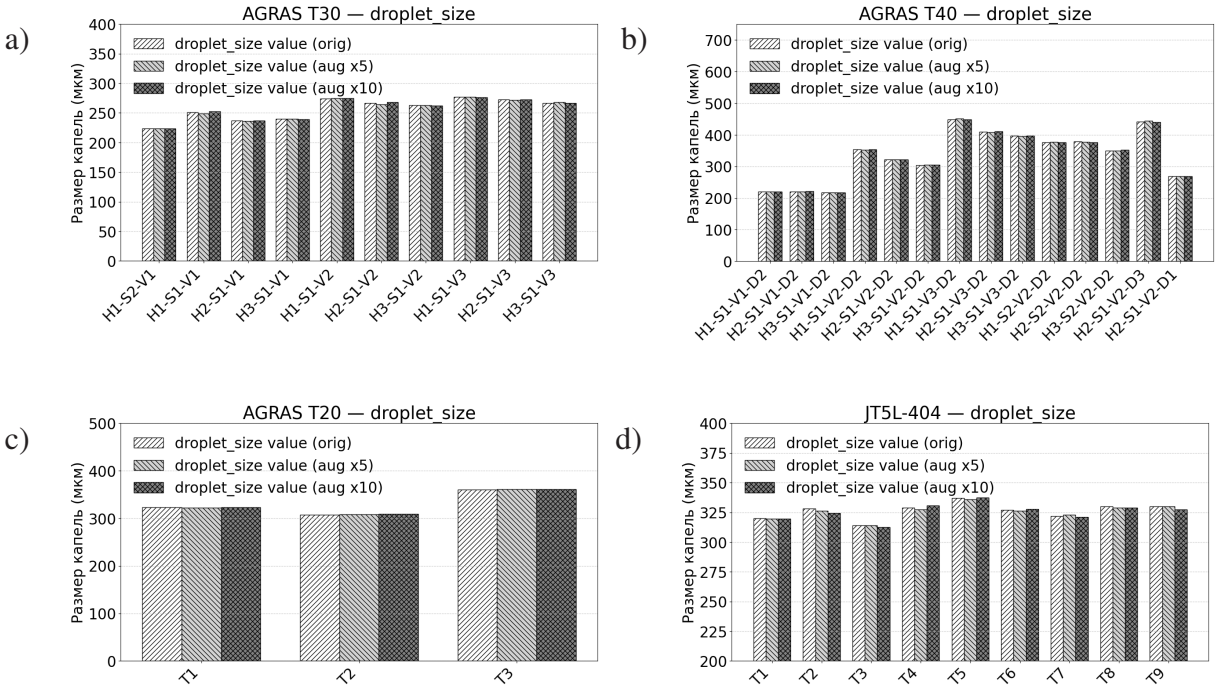

Рис.2.2. Распределение значений диаметра капель в исходной и аугментированных выборках: *a* — AGRAS T30; *b* — AGRAS T40; *c* — AGRAS T20; *d* — JT5L-404

По графикам на рис. 2.1 и 2.2 видно, что аугментация экспериментальных данных в пятикратном и десятикратном объёме не приводит к существенным изменениям в распределении целевых переменных. Во всех наборах наблюдается высокая согласованность между средними значениями покрытия и диаметра капель в оригинальных и расширенных выборках: столбцы для «orig», «aug ×5» и «aug ×10» располагаются без выраженных смещений.

2.2. Объединение и предварительная обработка данных

После завершения аугментации данные из трёх расширенных датасетов были объединены в один общий. На этом этапе выполнялось выравнивание названий признаков путём, например, разбиения параметров погодных характеристик (температура, влажность, скорость ветра) на минимальные и максимальные значения. При объединении отсутствующие столбцы автоматически заполнялись пустыми значениями (NaN) для тех записей, в которых соответствующие измерения не проводились.

Дальнейший анализ показал, что многие признаки — в основном связанные с морфологическими характеристиками растений и отдельными техническими параметрами БПЛА — содержат пропуски в большей части строк, поскольку не во всех статьях приводился полный набор измерений. Для алгоритма CatBoost такая ситуация не требовала дополнительной обработки: пропуски оставлены, за исключением признака `experiment.params.atomization_diameter`, где NaN заменены на значение `off`, отражающее отсутствие автоматизации у модели. Кроме того, столбец `experiment.name`, по результатам первоначального анализа, вносил сильную линейную зависимость между наблюдениями и целевыми переменными, поэтому он был удалён для исключения нежелательных корреляций.

В отличие от CatBoost, остальные алгоритмы не обладают встроенной поддержкой пропусков и требуют полного набора признаков. Поэтому все столбцы с хотя бы одним пустым значением были исключены, кроме тех, что непосредственно относятся к результатам эксперимента (значения покрытия и среднего диаметра капель), поскольку там пропуски означают лишь отсутствие измерений для одной из целевых метрик. После этого все оставшиеся категориальные признаки были преобразованы методом `one-hot encoding`, что позволило получить бинарные индикаторы для каждого уникального значения. Непрерывные числовые признаки

затем подверглись масштабированию, улучшая сходимость линейной регрессии и регрессии опорных векторов.

2.3. Выводы

Аугментация в пяти- и десятикратном объёме не искажает распределения ключевых метрик, что подтверждено визуальными сравнениями. Объединение и очистка данных обеспечили единую структуру: для CatBoost сохранены все исходные признаки с NaN, для остальных ансамблей исключены колонки с пропусками и применено one-hot кодирование, а для MLR/SVR к этому добавлено масштабирование числовых признаков. Таким образом, были получены три полноценных набора данных, оптимизированных под требования соответствующих алгоритмов и готовых к оценке.

ГЛАВА 3. СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ

В главе 3 проводится сравнительный анализ шести моделей, обученных на подготовленных выборках. Цель — оценить качество предсказания двух ключевых агрономических характеристик: площади покрытия и размера капель. Для обеих задач модели обучались на 80% данных. Оставшиеся 20% использовались для тестирования. Результаты сравниваются как по визуальному соответствию предсказаний реальным значениям, так и по метрикам качества — коэффициенту детерминации R^2 (1.29) и среднеквадратичной ошибке $RMSE$ (1.30). Также проводится выявление важнейших признаков, которые внесли наибольший вклад в предсказание целевых переменных.

3.1. Предсказание площади покрытия

3.1.1. Сравнение результатов предсказаний

На рис.3.1 видно, что линейная регрессия не способна точно аппроксимировать зависимости между признаками и покрытием. Предсказания на тестовой выборке значительно отклоняются от истинных значений. Это подтверждается метриками: $R^2 = 0.759$ и $RMSE = 1.29$. Полученные результаты указывают на несоответствие простой линейной модели характеру зависимости в данных, что ожидаемо при высокой размерности пространства признаков.

SVR (см. рис.3.2), использованный с подобранными параметрами методом поиска на сетке с оценкой через кросс-валидацию, показывает более выраженное приближение на обучении, однако на тестовой выборке наблюдается ухудшение качества предсказаний ($R^2 = 0.699$, $RMSE = 1.44$). Это отражает трудности SVR при работе с разнородными и высокоразмерными данными, несмотря на предварительный подбор гиперпараметров.

Случайный лес (см. рис.3.3) обеспечивает более высокую точность на обучающей выборке ($R^2 = 0.995$), но на тесте его результат уступает градиентным бустингам, демонстрируя $R^2 = 0.782$ при $RMSE = 1.23$. Несмотря на это, модель остаётся устойчивой.

Градиентный бустинг (см. рис.3.4) проявил себя надёжно: он уступает CatBoost, но сохраняет высокое качество на тесте ($R^2 = 0.815$, $RMSE = 1.13$), демонстрируя адекватную способность к обобщению сложных зависимостей.

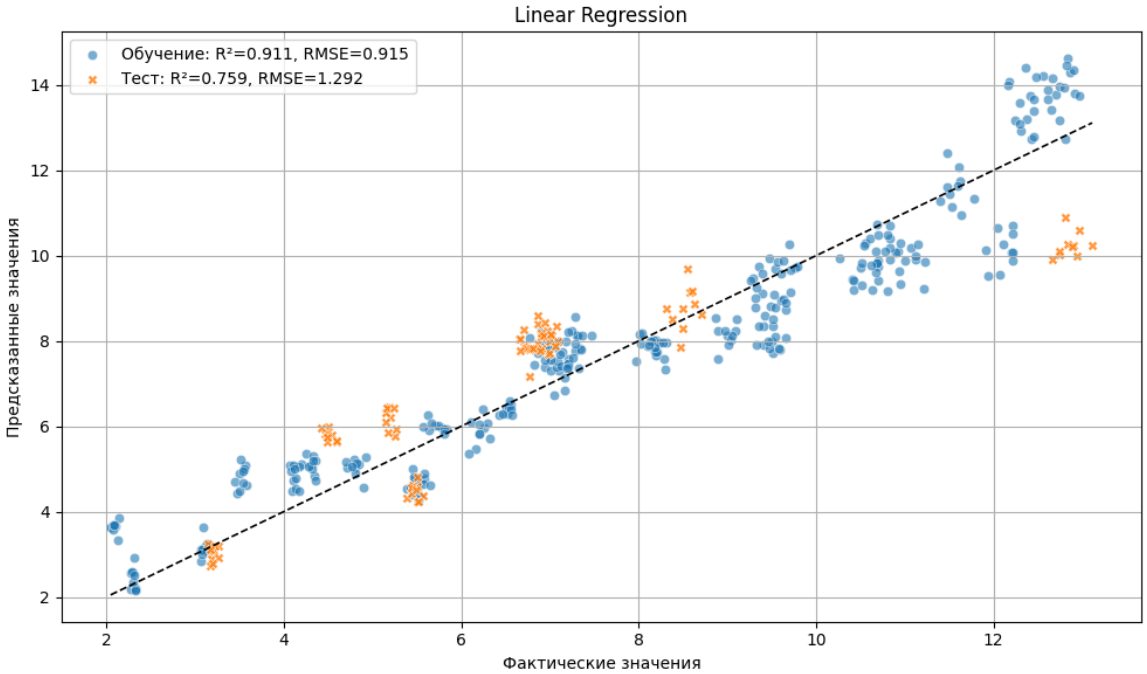

Рис.3.1. Предсказания линейной регрессии (LR)

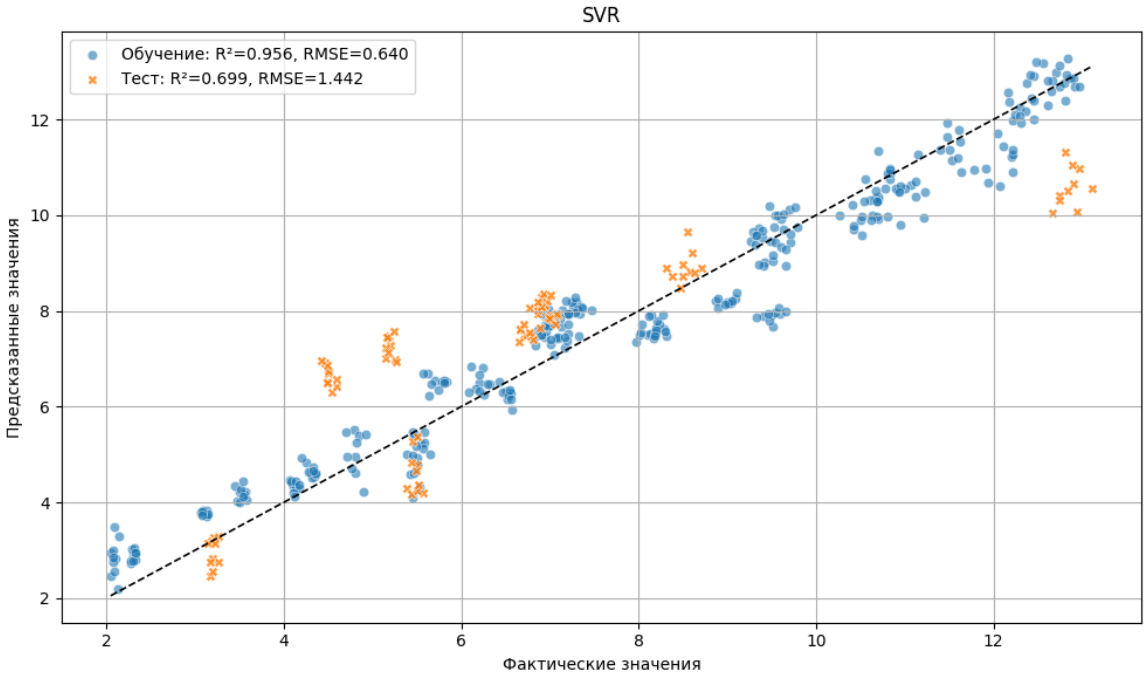

Рис.3.2. Предсказания метода опорных векторов (SVR)

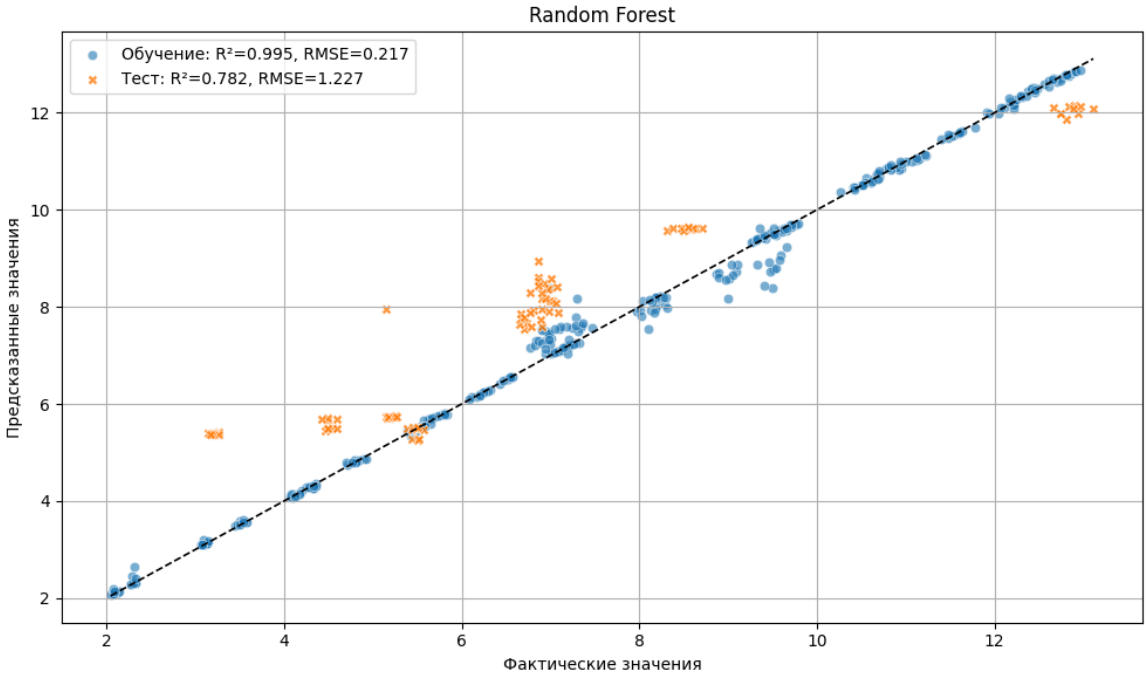

Рис.3.3. Предсказания метода случайного леса (RF)

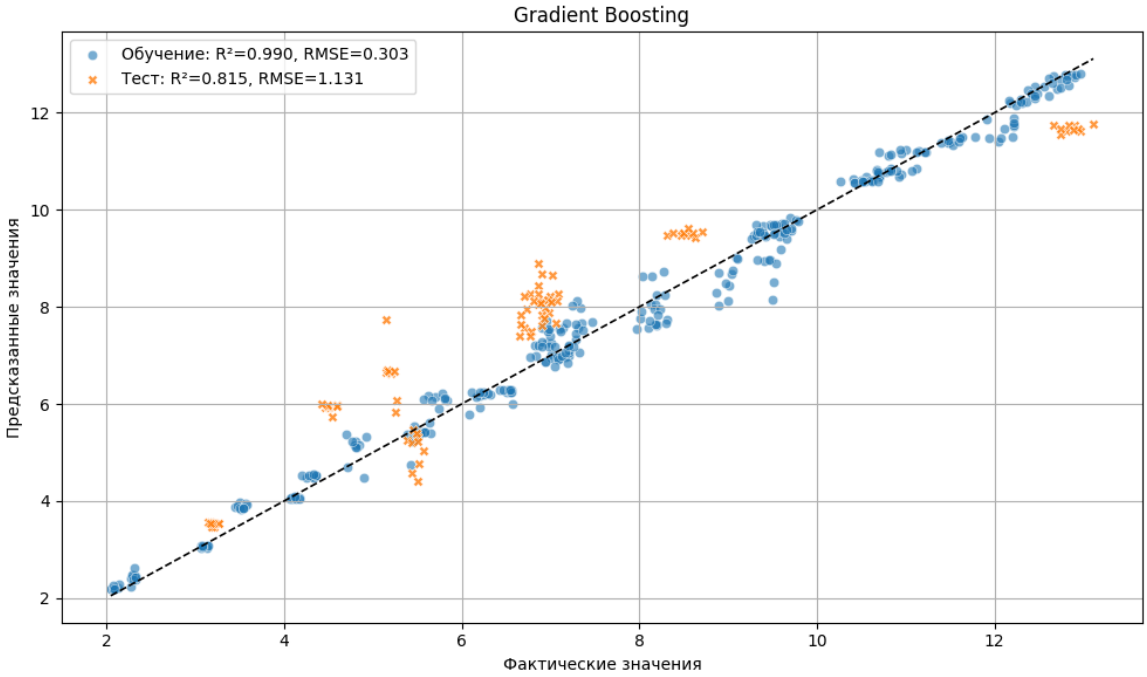

Рис.3.4. Предсказания метода градиентного бустинга (GBR)

Использование XGBoost с параметрами по умолчанию приводит к переобучению, что видно по практически идеальному обучающему $R^2 = 0.998$ и заметному падению точности на тесте ($R^2 = 0.793$). Это подчёркивается визуальной расфокусировкой предсказаний на графике (см. рис.3.5).

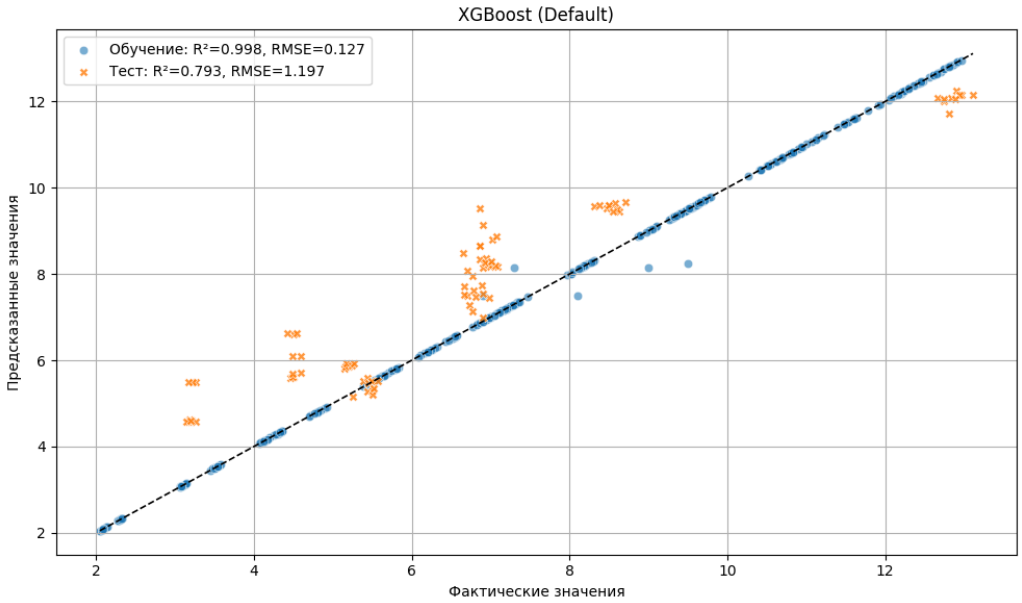

Рис.3.5. Предсказания XGBoost до настройки параметров

После настройки с помощью генетического алгоритма модель XGBoost частично смягчила переобучение (см. рис.3.6): качество на обучении снизилось, но улучшения на тесте практически не наблюдается ($R^2 = 0.792$, $RMSE = 1.20$).

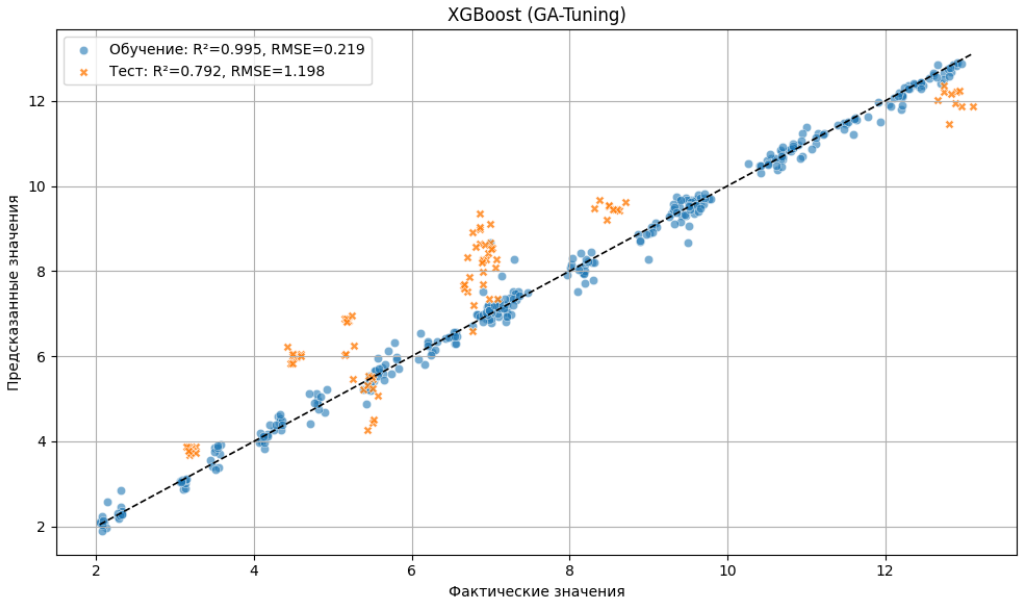

Рис.3.6. Предсказания XGBoost после настройки параметров

CatBoost продемонстрировал наилучший результат среди всех моделей (см. рис.3.7). Коэффициент детерминации на тестовой выборке составил $R^2 = 0.882$ при $RMSE = 0.90$. Это свидетельствует о высокой устойчивости модели к структурным особенностям данных и сильной способности к обобщению.

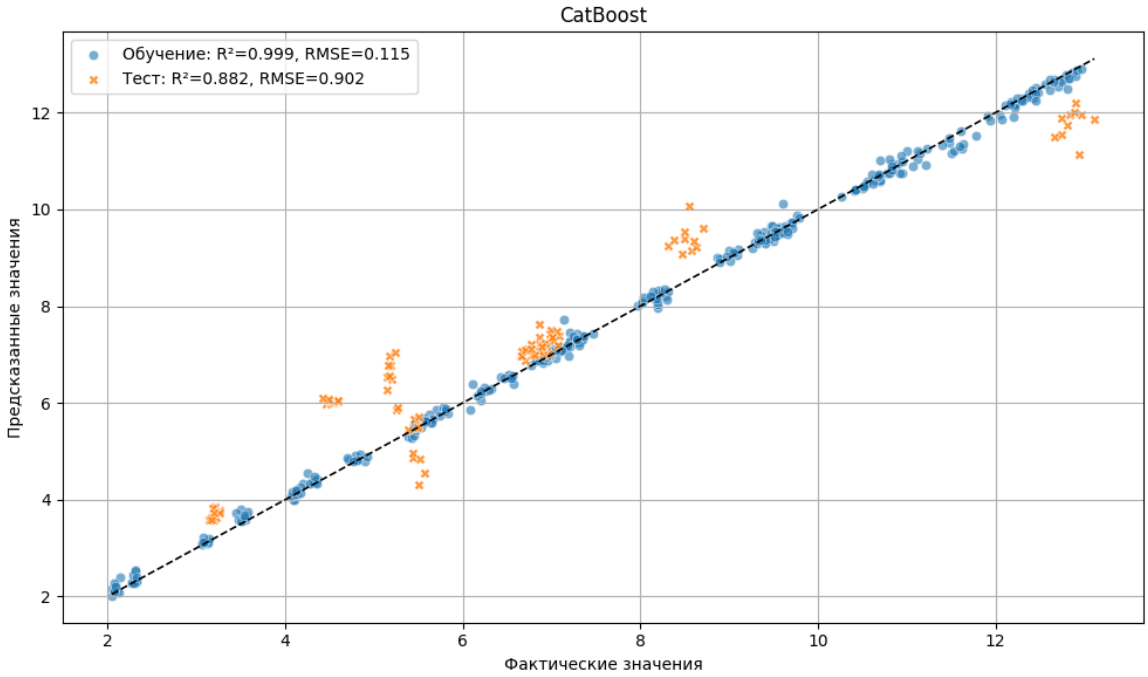

Рис.3.7. Предсказания модели CatBoost

В табл.3.1 отражены значения метрик для каждой модели:

Таблица 3.1

Сводная таблица R^2 и $RMSE$ для площади покрытия

Модель	Обучающая выборка		Тестовая выборка	
	R^2	$RMSE$	R^2	$RMSE$
CatBoost	0.998585	0.115213	0.882371	0.901760
GradientBoosting	0.990218	0.302911	0.814927	1.131111
XGBoost (Default)	0.998290	0.126628	0.792790	1.196849
XGBoost (GA-Tuning)	0.994903	0.218646	0.792265	1.198363
RandomForest	0.994976	0.217081	0.782113	1.227295
LinearRegression	0.910811	0.914635	0.758507	1.292069
SVR	0.956363	0.639769	0.699309	1.441762

Они подтверждают высокую устойчивость ансамблевых методов, особенно CatBoost, который демонстрирует наилучшее сочетание точности и обобщающей способности: самый высокий R^2 и минимальный $RMSE$ на тестовой выборке. GBR и RF также показывают высокую обучаемость, однако уступают CatBoost по точности на тестовых данных. Модели XGBoost, как в базовой конфигурации, так и после

настройки, демонстрируют признаки переобучения. LR и SVR являются наименее эффективными в условиях высокой размерности пространства признаков.

3.1.2. Выявление важных признаков

Рассмотрим, какие признаки оказались наиболее важными для прогнозирования площади покрытия в моделях, показавших лучшие результаты. В модели CatBoost объём распыления даёт 65.5% суммарной важности, скорость полёта — 8.0%, минимальная влажность воздуха — 3.4%, а число форсунок — 2.1%, тогда как вклад остальных факторов, включая конструктивные характеристики аппарата и оптимальные значения высоты полёта, не превышает 2%. Соответственно, интенсивность распыления и характеристики окружающего воздуха непосредственно определяют равномерность и площадь покрытия, что согласуется с выводами о том, что высокий объём распыления обеспечивает необходимую плотность осаждения, а скорость влияет на дрейф и покрытие кроны растений [4].

В модели GBR: объём распыления — 76.9%, минимальная температура воздуха — 9.4%, скорость полёта — 6.8%, остальные параметры — не более 1%.

В модели XGBoost: объём распыления — 57.6%, количество форсунок — 18.5%, мелкодисперсный режим атомизации — 7.5%, минимальная температура — 6.4%, объём бака — 5.8%, скорость полёта — 1.7%, остальные параметры — не более 0.5%.

Результаты указывают на то, что необходимо учитывать значимость параметров полёта и метеоусловий для повышения устойчивости прогноза. Кроме того, сформулированный вывод подтверждается работами, использующими компьютерную гидродинамику для моделирования аэродинамики ротора и распределения капель, где скорость полёта, высота и тип форсунок были определены как ключевые факторы равномерного нанесения [3], а эмпирические исследования демонстрируют, что параметры распыления и их калибровка существенно влияют на равномерность и эффективность обработки культур [6]. Таким образом, анализ важности признаков для прогнозирования площади покрытия подтвердил, что основными признаками выступают эксплуатационные параметры распыления и метеоусловия.

3.2. Предсказание диаметра капель

3.2.1. Сравнение результатов предсказаний

Как и в задаче покрытия, линейная регрессия не смогла качественно справиться с прогнозированием размера капель. На рис.3.8 видно сильное отклонение предсказаний от действительных значений. Метрики подтверждают низкую точность модели ($R = 0.660$, $RMSE = 30.70$).

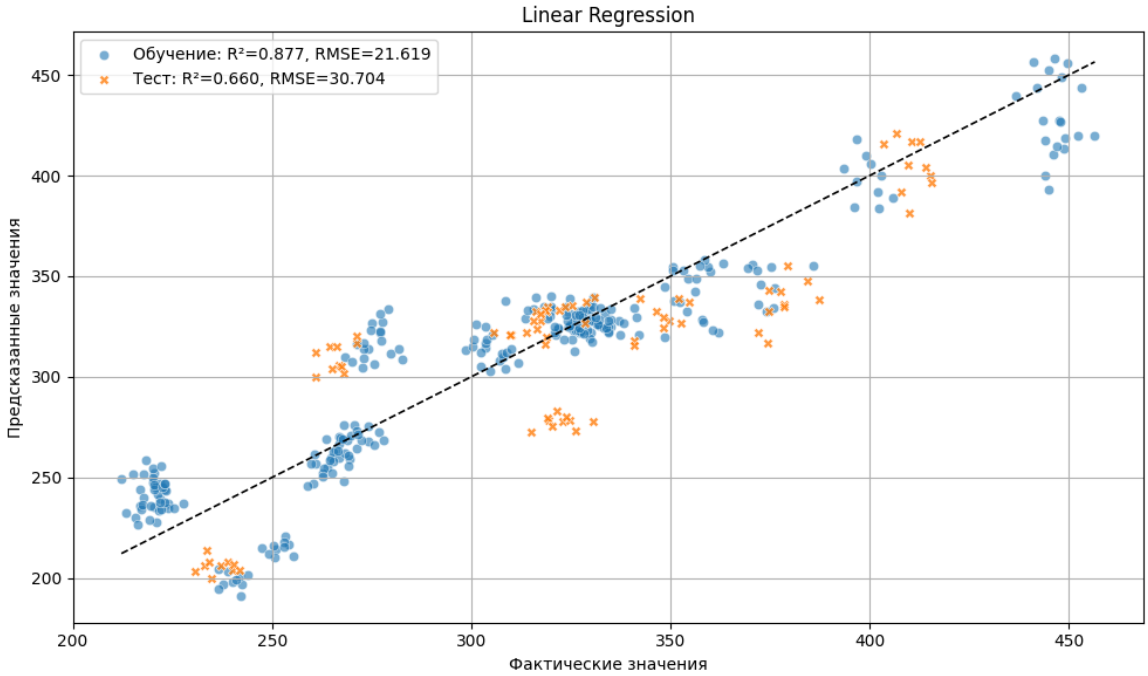

Рис.3.8. Предсказания линейной регрессии (LR)

SVR, несмотря на свою универсальность, также демонстрирует ограниченную обобщающую способность в этой задаче (см. рис.3.9). $RMSE$ на тесте приблизительно равен 22.96, а точность составляет лишь 0.810, что сопоставимо с результатами ансамблевых методов.

Случайный лес (см. рис.3.10) и градиентный бустинг (см. рис.3.11) обеспечили почти схожие результаты ($R^2 = 0.810$, $RMSE = 22.94$ и $R^2 = 0.820$, $RMSE = 22.35$ соответственно), при этом оба метода продемонстрировали неплохую аппроксимацию на обучающей выборке, но незначительное падение точности на тесте.

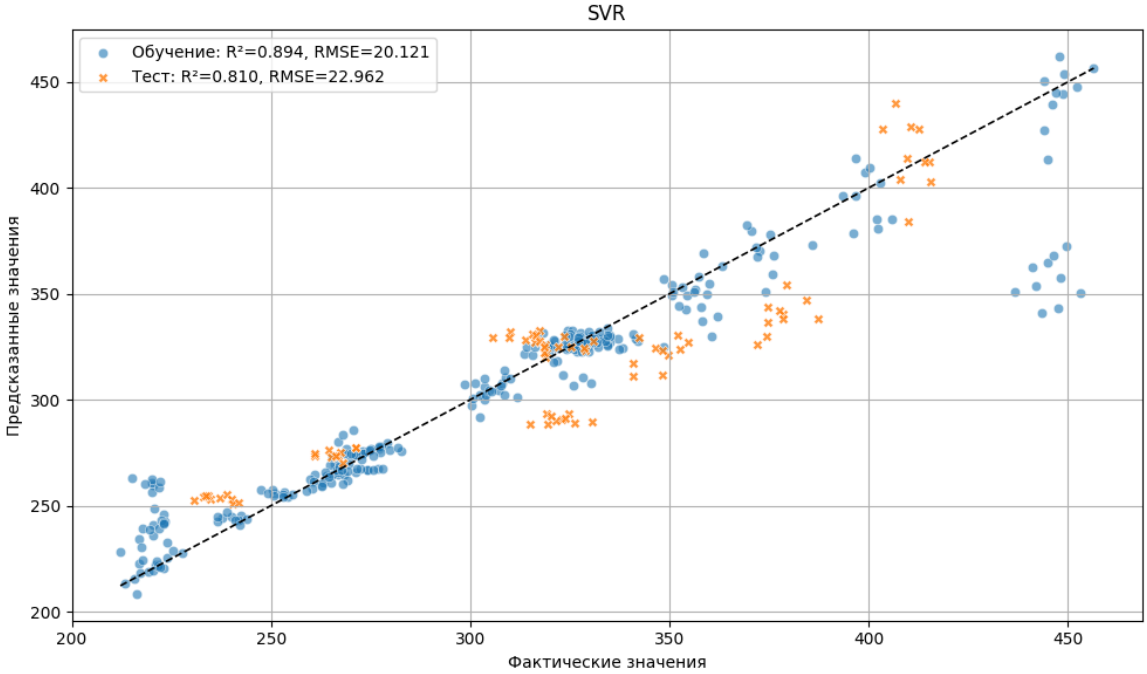

Рис.3.9. Предсказания метода опорных векторов (SVR)

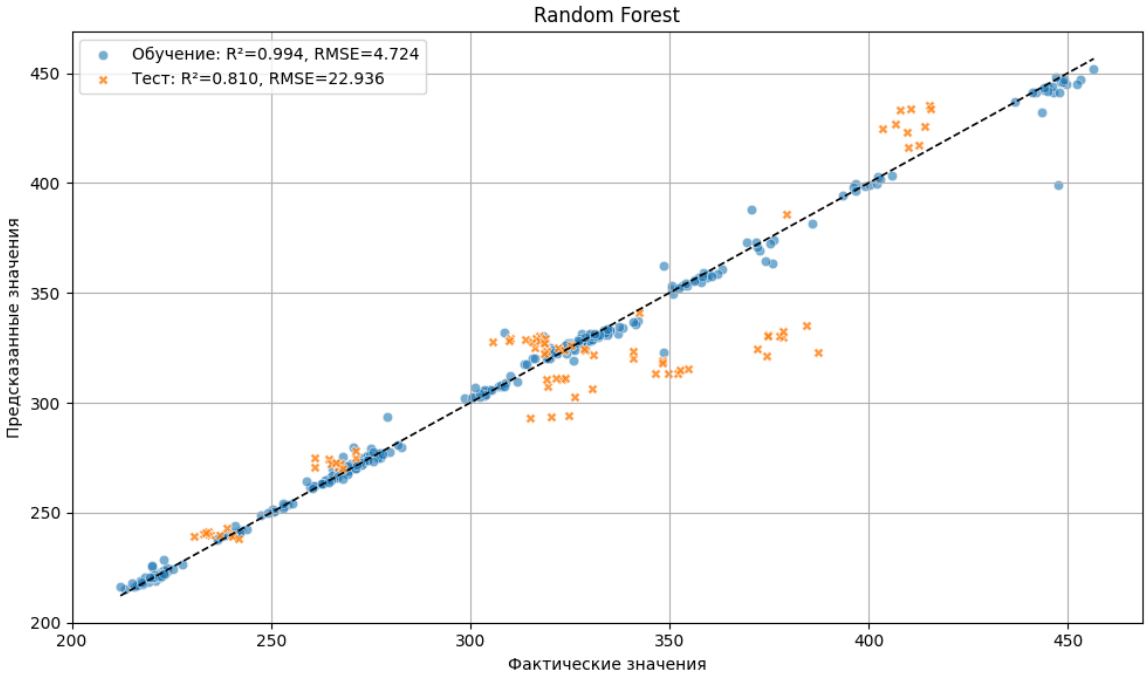

Рис.3.10. Предсказания метода случайного леса (RF)

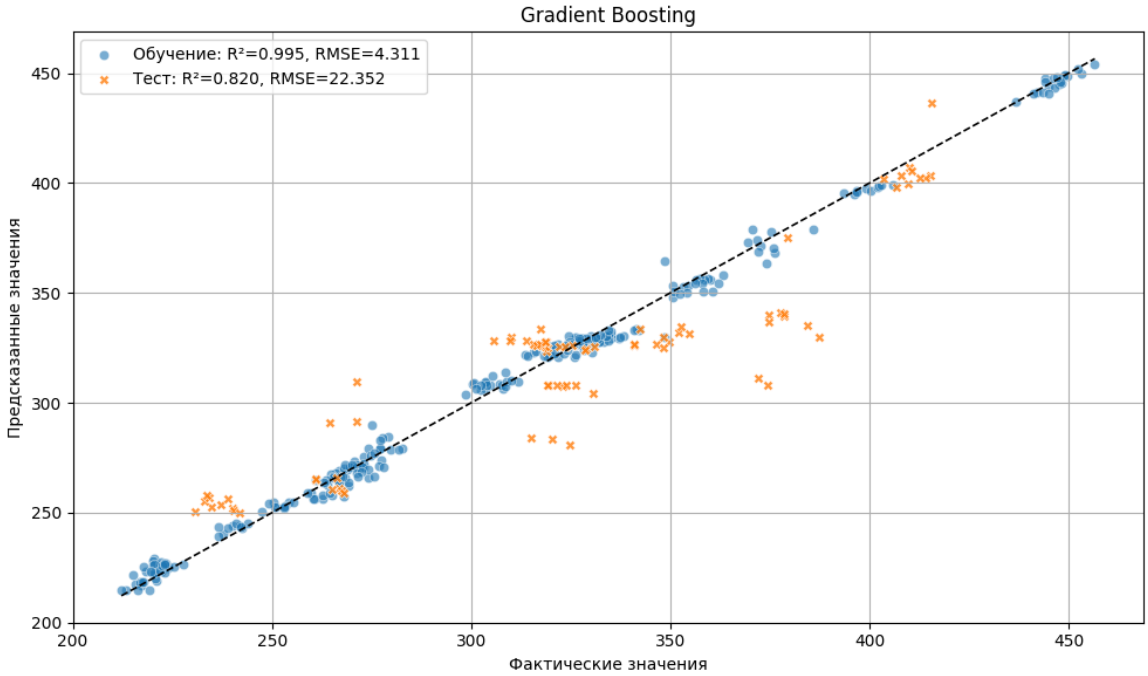

Рис.3.11. Предсказания метода градиентного бустинга (GBR)

Для задачи предсказания диаметра капель XGBoost также подвержен переобучению. На рис.3.12 видно почти полное совпадение предсказаний на обучении ($R^2 = 0.9999$, $RMSE = 0.57$), но значительное снижение точности на тесте ($R^2 = 0.863$, $RMSE = 19.47$). Попытка настройки XGBoost с помощью генетического алгоритма (см. рис.3.13) позволила несколько сгладить переобучение, однако общие показатели на тесте ухудшились ($R^2 = 0.811$, $RMSE = 22.90$).

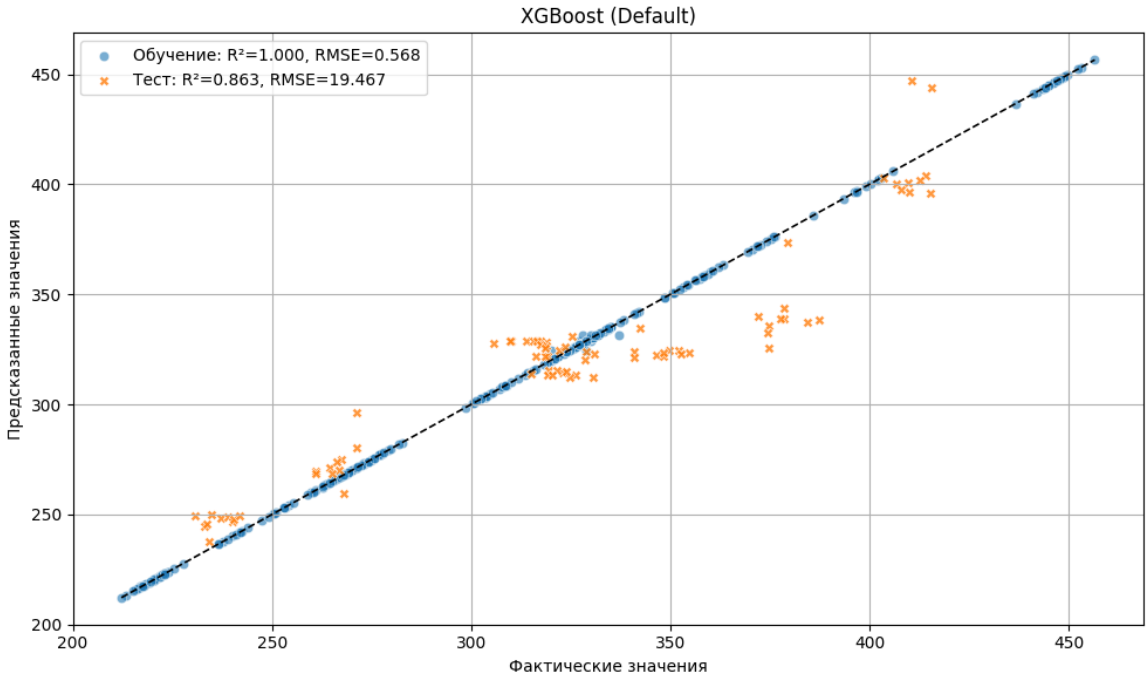

Рис.3.12. Предсказания XGBoost до настройки параметров

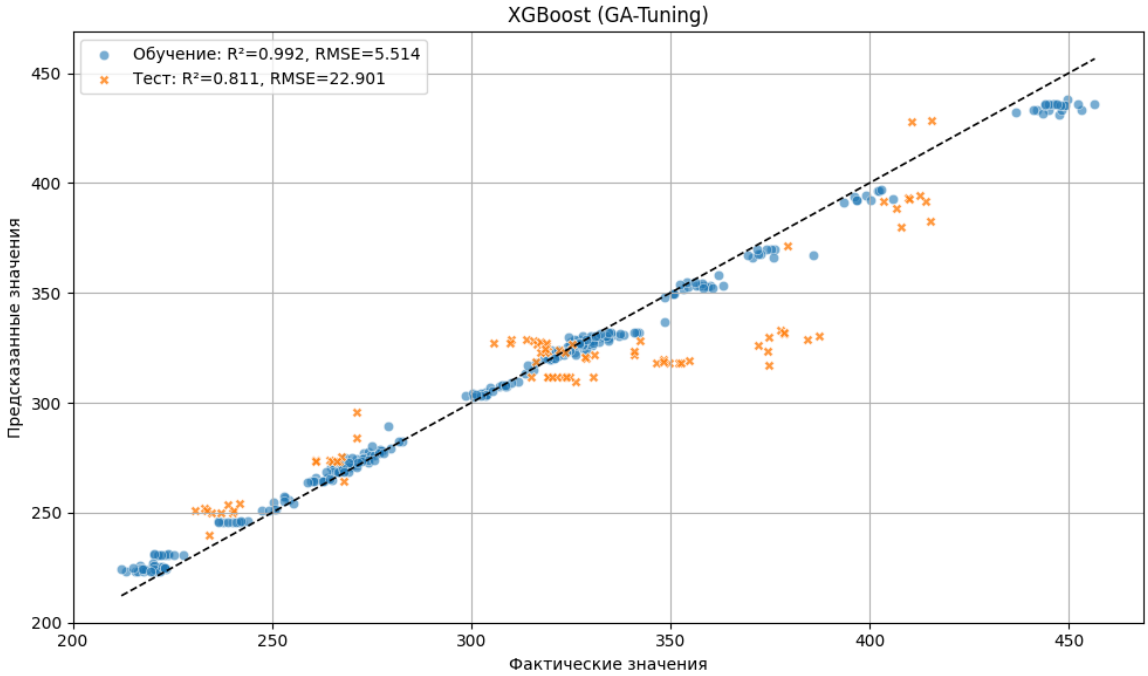

Рис.3.13. Предсказания XGBoost после настройки параметров

CatBoost снова показал лучшие результаты, обеспечив наилучшее соотношение точности и ошибки на тестовой выборке (см. рис.3.14). При $R^2 = 0.864$ и $RMSE = 19.44$ модель демонстрирует стабильную производительность и умеренное переобучение, что делает её предпочтительным выбором.

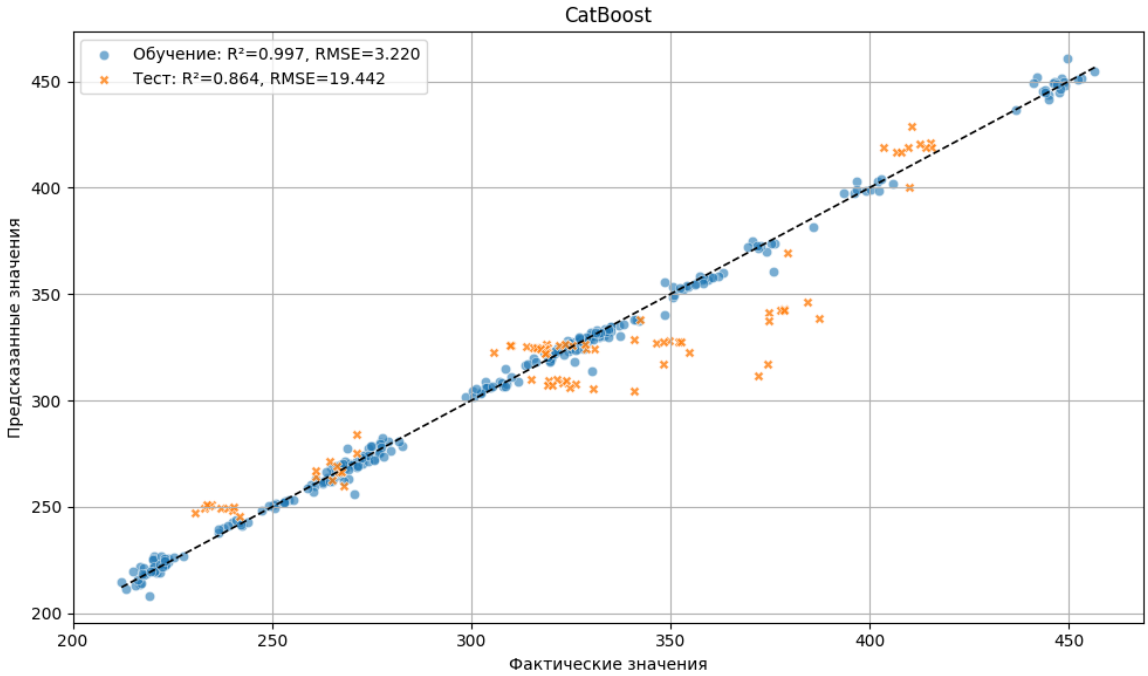

Рис.3.14. Предсказания модели CatBoost

В табл.3.2 сведены все значения метрик, полученные при предсказании диаметров капель:

Таблица 3.2

Сводная таблица R^2 и $RMSE$ для диаметра капель

Модель	Обучающая выборка		Тестовая выборка	
	R^2	$RMSE$	R^2	$RMSE$
CatBoost	0.997282	3.219687	0.863812	19.442279
XGBoost (Default)	0.999915	0.568252	0.863462	19.467204
GradientBoosting	0.995127	4.311346	0.820001	22.351813
XGBoost (GA-Tuning)	0.992029	5.513798	0.811052	22.900683
RandomForest	0.994150	4.723604	0.810464	22.936303
SVR	0.893861	20.120588	0.810034	22.962318
LinearRegression	0.877465	21.618812	0.660338	30.704413

Значения подтверждают, что наилучший результат обеспечил метод CatBoost. RF и GBR показали схожее качество на обучении, однако уступили на тесте. Модель XGBoost в базовой конфигурации достигла высокой точности, но сопровождалась резким ростом переобучения, которое лишь частично удалось смягчить с помощью настройки гиперпараметров. SVR и LR, несмотря на умеренную точность на обучающей выборке, продемонстрировали худшие результаты на тесте.

3.2.2. Выявление важных признаков

Теперь аналогично рассмотрим задачу прогнозирования диаметра капель. В CatBoost объём распыления составляет 55.4% общей важности, диаметр атомизации — 13.3%, высота полёта — 4.4%, мощность пропеллеров — 4.0%, скорость полёта — 2.7%, а количество форсунок — 2.4%, что указывает на прямую зависимость размеров капель от объёма подачи, настроек распылительного потока и параметров полёта. Важность остальных признаков не превышает 2%.

При этом в XGBoost на первом месте по важности — мелкодисперсный режим атомизации (35.4%), за ним следуют объём бака (25.7%) и крупнодисперсный режим (24.3%), а объём распыления здесь составляет лишь 13.0%, что подчёркивает роль точной настройки форсунок для формирования требуемого спектра капель [3; 6]. Остальные признаки составили не более 0.6% важности.

В модели GBR объём распыления достигает 57.1%, крупнодисперсный режим атомизации — 10.5%, минимальная скорость ветра — 7.6%, мелкодисперсный режим — 6.2%, объём бака — 5.7%, высота полёта — 3.3%, скорость полёта — 2.2%, остальные признаки — не более 1.9%. Это подтверждает сильное влияние как настроек форсунок, так и метеоусловий на размер капель и их траекторию [3].

Для задачи прогнозирования диаметра капель самыми важными оказались объём и режимы распыления, хотя структура их вкладов несколько отличается от задачи покрытия. Также не малый вклад внесли аэродинамические факторы.

3.3. Выводы

Проведённый анализ продемонстрировал устойчивое преимущество ансамблевых моделей градиентного бустинга над линейными и простыми нелинейными подходами при решении задач регрессии в условиях агротехнических данных. CatBoost показал наилучшее сочетание точности и устойчивости в обеих задачах, подтверждая свою пригодность для работы с неоднородными и неполными данными. XGBoost, несмотря на высокую точность на обучении, склонен к переобучению, что требует дополнительной настройки или регуляризации. Модели без бустинга, линейная регрессия и SVR, существенно уступают как по точности, так и по устойчивости.

Обобщая результаты важности признаков, можно заключить, что для обеих задач — как площади покрытия, так и диаметра капель — параметры объёма распыления и скорость полёта остаются ключевыми. При прогнозировании диаметра капель особое значение приобретает режим атомизации. Также немаловажными оказались число форсунок и объём бака.

ЗАКЛЮЧЕНИЕ

Современные методы машинного обучения предоставляют инструментарий, позволяющий не только сократить объём полевых исследований, но и повысить точность их оценок. Особенно актуальным становится построение предсказательных моделей на основе агрегированных данных, полученных из открытых источников, включая научные публикации, а также их последующая проверка на предмет пригодности к решению прикладных задач агроинженерии.

В ходе выполнения ВКР была достигнута главная цель - обучение и сравнение регрессионных моделей, предназначенных для прогнозирования эффективности распыления инсектицида с применением БПЛА. При этом эффективность характеризуется площадью покрытия (%) и диаметром капель (мкм) инсектицида. Для достижения цели были проделаны следующие этапы.

Во-первых, сформирован обучающий датасет, агрегирующий ключевые значения эффективности распыления, полученные из различных независимых источников в открытом доступе. Для повышения полноты и устойчивости выборки была применена процедура аугментации: в частности, использованы методы генерации симметричных распределений и масштабирования.

Во-вторых, на подготовленном датасете были обучены шесть моделей регрессии: множественная линейная регрессия (MLR), регрессия опорных векторов (SVR), метод случайного леса (RF), градиентный бустинг (GBR), XGBoost и CatBoost. Все модели были протестированы на двух ключевых задачах — предсказание площади покрытия и диаметра капель.

В-третьих, проведён сравнительный анализ точности построенных моделей. Полученные результаты показали, что CatBoost обеспечивает наивысшую обобщающую способность при стабильных метриках как на обучении, так и на тестировании. Метод случайного леса и градиентный бустинг показали хорошие, но менее устойчивые результаты, чувствительные к структуре данных. XGBoost проявил признаки переобучения, а линейная регрессия в обоих случаях оказалась неконкурентоспособной, что указывает на нелинейные зависимости в исходных данных. SVR также продемонстрировал низкую эффективность, даже при настройке гиперпараметров.

В-четвёртых, интерпретация результатов с точки зрения значимости входных факторов позволила выявить ключевые признаки, влияющие на площадь покрытия и диаметр капель инсектицида. Для обоих задач особую важность предоставили

объём распыления и скорость полёта. Однако для прогнозирования площади покрытия немаловажными оказались метеорологические условия, а для диаметра капель — режим атомизации и технические характеристики БПЛА.

Таким образом, можно сделать общий вывод: агрегированные и аугментированные данные, собранные из открытых источников, обладают высокой прогностической ценностью при условии правильной предварительной обработки. Построенные на их основе модели машинного обучения продемонстрировали высокую точность и применимость к реальным задачам оценки параметров распыления при использовании БПЛА.

На основании полученных результатов можно предложить следующие рекомендации:

A. При проектировании агрономических экспериментов с использованием БПЛА целесообразно учитывать возможность последующего машинного анализа данных. Для этого важно стандартизировать структуру и формат параметров.
B. Модель CatBoost рекомендуется использовать в качестве базовой для задач предсказания покрытия и диаметра капель, с последующей тонкой настройкой под конкретные условия.
C. Для снижения чувствительности к выбросам рекомендуется применять методы предварительной фильтрации данных либо адаптивного взвешивания ошибок при обучении.
D. Системы поддержки агротехнологических решений могут быть дополнены предсказательными модулями, построенными на основе моделей машинного обучения, обученных на объединённых выборках с привлечением открытых научных источников.

В заключение следует подчеркнуть, что развитие цифровых подходов в агрономии, включая машинное обучение, открывает новые возможности для оптимизации процессов защиты растений, повышения эффективности агротехнологий и внедрения точного земледелия. Полученные в данной работе результаты подтверждают практическую применимость современных методов анализа данных в агроинженерной практике и могут служить основой для дальнейших исследований и разработок в этой области.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Параметры применения беспилотных летательных аппаратов при обработке средствами защиты растений сельскохозяйственных культур / Н. Ю. Курченко [и др.] // Известия НВ АУК. — 2023. — Т. 69, № 1. — С. 527—536. — DOI 10.32786/2071-9485-2023-01-58.

2. Research on the application of lightweight YOLO model in detection of small crop diseases and pests / Q. Yang [et al.] // Journal of Chinese Agricultural Mechanization. — 2024. — Vol. 45, no. 9. — P. 265–270.

3. Optimization of Flight Mode and Coupling Analysis of Operational Parameters on Droplet Deposition and Drift of Unmanned Aerial Spraying Systems (UASS) / Q. Liu [et al.] // Agronomy. — 2025. — Vol. 15. — P. 367. — DOI 10.3390/agronomy15020367.

4. Optimizing Unmanned Aerial Vehicle Operational Parameters to Improve Pest Control Efficacy and Decrease Pesticide Dosage in Tea Gardens / M. Wu [et al.] // Agronomy. — 2025. — Vol. 15. — P. 431. — DOI 10.3390/agronomy15020431.

5. *Lopes L. d. L.*, *Cunha J. P. A. R. d.*, *Nomelini Q. S. S.* Use of Unmanned Aerial Vehicle for Pesticide Application in Soybean Crop // AgriEngineering. — 2023. — Vol. 5, no. 4. — P. 2049–2063. — DOI 10.3390/agriengineering5040126.

6. Effect of flight operative height and genotypes on conilon coffee spraying using an unmanned aerial vehicle / E. L. d. Vitória [et al.] // Coffee Science. — 2022. — Vol. 17. — e172003. — DOI 10.25186/.v17i.2003.

7. UAV-spray application in vineyards: Flight modes and spray system adjustment effects on canopy deposit, coverage, and off-target losses / A. Biglia [et al.] // Science of The Total Environment. — 2022. — Vol. 845. — P. 157292. — DOI 10.1016/j.scitotenv.2022.157292.

8. Machine Learning in Agriculture: A Review / K. G. Liakos [et al.] // Sensors. — 2018. — Vol. 18, no. 8. — DOI 10.3390/s18082674.

9. *Jordan M. I.*, *Mitchell T. M.* Machine learning: Trends, perspectives, and prospects // Science. — 2015. — Vol. 349, no. 6245. — P. 255–260. — DOI 10.1126/science.aaa8415.

10. *Rao C. R.* Linear Statistical Inference and Its Applications. — 2nd ed. — New York: Wiley, 1973.

11. *Vapnik V. N.* The Nature of Statistical Learning Theory. — New York: Springer, 1995. — DOI 10.1007/978-1-4757-2440-0.

12. Support vector regression machines / H. Drucker [et al.] // Adv Neural Inform Process Syst. — 1997. — Vol. 28. — P. 779–784.

13. *Breiman L.* Random Forests // Machine Learning. — 2001. — Vol. 45, no. 1. — P. 5–32. — DOI 10.1023/A:1010933404324.

14. *Friedman J. H.* Greedy Function Approximation: A Gradient Boosting Machine // Annals of Statistics. — 2001. — Vol. 29, no. 5. — P. 1189–1232. — DOI 10.1214/aos/1013203451.

15. *Hastie T.*, *Tibshirani R.*, *Friedman J.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — 2nd ed. — New York: Springer, 2009.

16. *Chen T.*, *Guestrin C.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — San Francisco, California, USA: Association for Computing Machinery, 2016. — P. 785–794. — (Ser.: KDD '16). — DOI 10.1145/2939672.2939785.

17. CatBoost: Unbiased Boosting with Categorical Features / L. Prokhorenkova [et al.] // Advances in Neural Information Processing Systems. Vol. 31. — Curran Associates, Inc., 2018. — P. 6638–6648. — URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf (visited on 24.05.2025).

18. *Fisher R. A.* The goodness of fit of regression formulae, and the distribution of regression coefficients // Journal of the Royal Statistical Society. — 1922. — Vol. 85, no. 4. — P. 597–612. — DOI 10.2307/2340521.

19. *Mitchell M.* An Introduction to Genetic Algorithms // MIT Press. — 1998.

20. *Shorten C.*, *Khoshgoftaar T. M.* A survey on image data augmentation for deep learning // Journal of Big Data. — 2019. — Т. 6, № 1. — С. 60. — DOI 10.1186/s40537-019-0197-0.

21. *Feng C.*, *Yan D.*, *Hong T.* A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data // Energy and Buildings. — 2021. — Т. 231. — С. 110379. — DOI 10.1016/j.enbuild.2020.110379.