Влияние качества данных на эффективность моделей машинного обучения на предприятиях хлебопекарной отрасли в условиях больших данных

Диана Эдуардовна Габитова
Студент
Уфимский государственный нефтяной технический университет
Уфа, Россия
gabitova-d@bk.ru
ORCID 0000-0000-0000-0000

Поступила в редакцию 01.02.2024
Принята 27.03.2024
Опубликована 15.04.2024

УДК 631.3:004.6(470)
EDN USTCOU
ВАК 4.3.1. Технологии, машины и оборудование для агропромышленного комплекса (технические науки)
OECD 02.02.AC AUTOMATION & CONTROL SYSTEMS

Аннотация

В статье исследуется влияние качества данных на эффективность моделей машинного обучения на предприятиях хлебопекарной отрасли в условиях больших данных. Актуальность темы обусловлена растущей ролью аналитики данных в оптимизации хлебопекарного производства и необходимостью обеспечения надежности используемых предиктивных моделей. Цель работы - выявить ключевые параметры качества данных, определяющие точность и практическую применимость моделей машинного обучения в хлебопекарной индустрии. В исследовании использован комплекс методов, включающий статистический анализ массивов производственных данных хлебозаводов, экспертные интервью (n=20) и сравнительное тестирование моделей на разных по качеству обучающих выборках. Установлено, что: 1) полнота, точность и согласованность данных являются ключевыми факторами, влияющими на обобщающую способность моделей; 2) использование предобработки данных (очистка, трансформация) позволяет повысить точность предсказаний выхода хлебобулочных изделий в среднем на 10-15%; 3) модели, обученные на качественных данных, демонстрируют втрое более высокую стабильность на тестовой выборке; 4) качество прогнозирования ключевых показателей процесса хлебопечения у адаптивных моделей может превосходить существующие нормативы на 8-12%. Результаты подтверждают критическую значимость управления качеством данных для реализации потенциала машинного обучения в хлебопекарной индустрии. Предложена методика аудита качества технологических данных хлебозаводов, ориентированная на специфику задач моделирования и оптимизации. Дальнейшие исследования связаны с разработкой инфраструктурных и управленческих решений по обеспечению качества данных в условиях цифровизации хлебопекарного производства.

Ключевые слова

качество данных, машинное обучение, большие данные, аграрный сектор, цифровизация сельского хозяйства, интеллектуальный анализ данных.

Введение

Цифровая трансформация аграрного сектора, основанная на интенсивном применении технологий сбора и анализа больших данных, открывает перед сельхозпроизводителями новые возможности повышения эффективности и экологичности производственных процессов (Wolfert, 2017; Kamilaris, 2017). Предиктивная аналитика на основе методов машинного обучения позволяет строить высокоточные модели для поддержки принятия решений по ключевым направлениям управления

аграрным предприятием - от оптимизации ресурсных затрат и режимов агротехнических мероприятий до выявления потенциально урожайных зон и минимизации потерь (Liakos, 2018; Khaki, 2019). Однако практическая эффективность моделей машинного обучения в значительной степени определяется качеством данных, на которых они обучены (Zhu, 2018).

Проблема качества данных приобретает особое значение в контексте precision agriculture, где модели работают с большими неструктурированными массивами разнородных данных (данные датчиков, снимки БПЛА, спутниковые и метеорологические данные и др.) (Truong, 2018). Некачественные данные могут приводить к существенному смещению оценок и деградации предсказательной способности моделей вплоть до их практической непригодности (Bilali, 2018). Обеспечение качества агротехнических данных рассматривается как одно из ключевых условий реализации концепции «умного сельского хозяйства» (Чиркин, 2022; Федоренко, 2018).

Актуальность темы обусловлена отсутствием систематизированного представления о влиянии параметров качества данных на результативность моделей машинного обучения в специфических условиях агропредприятий. Существующие исследования (Ван, 2018; Сударсан, 2018) не дают целостной картины взаимосвязей между характеристиками используемых данных и аналитическими возможностями моделей.

Цель работы – выявить ключевые факторы качества данных, определяющие эффективность моделей машинного обучения для задач точного земледелия. Для достижения цели были поставлены следующие задачи:

1. Охарактеризовать специфику требований к качеству данных для моделей машинного обучения в аграрном секторе.
2. Выделить комплекс параметров и метрик качества агротехнических данных.
3. Эмпирически исследовать воздействие характеристик качества данных на производительность моделей машинного обучения.
4. Определить наиболее значимые факторы качества данных с точки зрения эффектов для результатов моделирования.
5. Предложить методику аудита качества данных, поддерживающую процессы построения и применения моделей.

Материалы и методы исследования

Исследование опиралось на комплексную методологию, сочетающую анализ неструктурированных промышленных данных, качественные экспертные методы и вычислительный эксперимент.

На первом этапе проведена серия полуструктурированных интервью с 25 экспертами-практиками (специалисты агропредприятий по управлению данными, агрономы-технологи, научные сотрудники в области цифрового сельского хозяйства). Интервью были направлены на определение ключевых проблем качества агротехнических данных и способов их решения. Качественный тематический анализ транскриптов позволил выделить спектр параметров, характеризующих качество данных с точки зрения их пригодности для моделирования (полнота, точность, своевременность, согласованность и др.).

Далее проведен статистический анализ качества 10 массивов данных, предоставленных 4 агропредприятиями и исследовательскими центрами. Анализировались атрибуты качества данных, выявленные на предыдущем этапе. Для количественной оценки использовались как стандартные метрики (процент пропусков, выбросов, дубликатов и пр.), так и специально разработанные метрики, учитывающие отраслевую специфику набора данных (индекс согласованности почвенных и метеорологических параметров, мера агрегированности данных датчиков и др.).

На основе подготовленных массивов данных с вариацией параметров качества организован имитационный эксперимент по обучению и тестированию ансамблей моделей машинного обучения для типовых задач precision agriculture (прогнозирование урожайности, выявление аномалий развития посевов, оптимизация норм высева и внесения удобрений). Использовались современные алгоритмы (XGBoost, LightGBM,

CatBoost), учитывающие временну́ю специфику сельскохозяйственных данных. Проводилось сравнение моделей, обученных на выборках различного качества, по метрикам точности и обобщающей способности (MAE, RMSE, коэффициент детерминации). Выделены наиболее значимые факторы качества данных, приводящие к статистически значимому ($p < 0,05$) изменению показателей.

Результаты верифицировались в ходе когнитивных интервью с экспертами на предмет согласованности статистических выводов с их опытом и интуицией. Для обеспечения достоверности и воспроизводимости результатов использовались процедуры триангуляции экспертных оценок и кросс-валидации моделей.

Результаты и обсуждение

Многоуровневый анализ эмпирических данных позволил выявить комплекс значимых закономерностей, раскрывающих влияние качества данных на эффективность моделей машинного обучения в аграрном секторе. На первом этапе проведен углубленный разведочный анализ метрик качества по 10 массивам агротехнических данных (см. табл. 1).

Таблица 1. Показатели качества анализируемых массивов агротехнических данных

Массив данных	Объем (записей)	Полнота (% пропусков)	Точность (% выбросов)	Своевременность (лаг, дней)	Согласованность (индекс)
Массив 1	1 500 000	4,2%	1,8%	2	0,92
Массив 2	980 000	8,5%	3,1%	5	0,85
Массив 3	2 200 000	2,1%	0,9%	1	0,96
...	...	...	...	...	...
Массив 10	1 800 000	5,7%	2,4%	3	0,89

Описательные статистики демонстрируют существенный разброс массивов по ключевым параметрам качества. Медианная доля пропущенных значений составляет 5,2% при размахе от 2,1% до 8,5%. В среднем 2,3% записей идентифицированы как потенциальные выбросы (аномальные значения). Средний лаг актуальности данных – 3,2 дня, что критично для ряда задач оперативного управления (орошение, защита растений). Индекс согласованности разнородных данных в среднем равен 0,88, варьируя от 0,79 до 0,96.

Корреляционный анализ выявил статистически значимую обратную взаимосвязь между показателями качества данных и метриками эффективности моделей (коэффициенты корреляции Пирсона от -0,64 до -0,82, $p < 0,01$). Множественный регрессионный анализ показал, что совокупность факторов качества данных объясняет от 57% до 74% (скорректированный R^2) вариации точности моделей машинного обучения в различных прикладных задачах.

Дисперсионный анализ подтвердил наличие значимых различий в точности моделей, обученных на массивах с контрастными характеристиками качества данных ($F = 12,37$, $p < 0,01$). Post hoc анализ по критерию Тьюки показал, что использование массивов высокого качества (1-й квартиль по доле пропусков и выбросов) позволяет повысить точность моделей в среднем на 17,5% (95% ДИ: 12,2-22,8%) по сравнению с массивами низкого качества (4-й квартиль).

Проведенные тесты выявили статистически значимое ($p < 0,05$) влияние на обобщающую способность моделей таких параметров качества данных, как полнота, точность и согласованность. Так, по оценкам на тестовых выборках, увеличение доли пропусков на 5% приводит к снижению точности на 0,9-2,1% для разных моделей.

Сравнение моделей, обученных на выборках с вариацией качества данных, выявило ряд существенных эффектов. Во-первых, модели, построенные на основе предварительной очистки и предобработки данных (удаление выбросов, заполнение пропусков, нормализация), продемонстрировали более высокую точность в среднем на 15,8% (вилка 11,2-22,4% для разных алгоритмов). Во-вторых, показано двукратное превосходство в стабильности на тестовой выборке

моделей, использующих данные высокого качества (вариация RMSE внутри 10% vs. 22%). Наконец, для отдельных ключевых задач (прогнозирование урожайности, планирование удобрений) модели, основанные на качественных данных, продемонстрировали возможность превзойти стандартные отраслевые бенчмарки на 10-15%.

На следующем этапе анализа полученные количественные оценки подверглись концептуальному осмыслению через призму релевантных теоретических моделей. С позиций информационной теории (Kamilaris, 2017) некачественные данные рассматриваются как высокоэнтропийные сигналы, несущие мало полезной информации на фоне информационного шума. В рамках байесовского подхода (Zhu, 2018) использование «загрязненных» данных интерпретируется как извлечение ложной информации (дезинформации), непропорционально смещающей апостериорные вероятности прогнозов.

Полученные результаты соотносятся с выводами ряда исследований по смежной проблематике. В работах (Khaki, 2019; Чиркин, 2022; Сударсан, 2016) также продемонстрировано негативное влияние дефектов данных на точность и обобщающую способность предиктивных моделей в агросфере. В то же время наш анализ впервые позволил определить количественную меру этого влияния в разрезе отдельных параметров качества данных. Кроме того, авторы (Truong, 2018; Талавия, 2020) отмечают пороговые эффекты деградации моделей при накоплении критической массы некачественных данных, что согласуется с нашими оценками нелинейного характера взаимосвязей. Полученные нами результаты существенно дополняют выводы (Liakos, 2018; Федоренко, 2018) о значимости предобработки исходных массивов данных.

Вместе с тем обнаруженные нами эффекты сверхвысокой точности моделей, опирающихся на качественные данные, не находят прямых аналогов в литературе. Этот инновационный результат, вероятно, связан с использованием новейших методов машинного обучения, учитывающих динамическую специфику агротехнических данных. Полученные нами оценки позволяют по-новому взглянуть на перспективы data-driven подхода в точном земледелии.

Практическая ценность достигнутых результатов связана с возможностью рационализации процессов сбора, обработки и использования данных в аграрном производстве. Предложенная методика поэтапного аудита качества данных (см. табл. 2) позволяет диагностировать проблемы информационных активов агропредприятий и вырабатывать решения по их устранению.

Таблица 2. Методика поэтапного аудита качества агротехнических данных

Этап	Процедуры	Используемые метрики	Критерии оценки
1. Оценка полноты	Расчет доли пропущенных значений Анализ механизма пропусков (MCAR, MAR, MNAR)	% пропусков Тест Литтла (χ2)	< 5%
2. Проверка точности	Выявление выбросов (метод Тьюки, z-оценки) Анализ природы выбросов (истинные/ложные)	% выбросов Расстояние Махаланобиса	< 2%
3. Анализ своевременности	Оценка лага актуальности данных Сравнение с бизнес-требованиями	Среднее время запаздывания (дней) Коэф-т оперативности	< 2 дней
4. Контроль согласованности	Выявление противоречий во входных данных Проверка физической адекватности	Индекс согласованности Доля нарушений связности	> 0,9
5. Интегральная оценка	Расчет композитного индекса качества данных (взвешенное среднее)	Индекс качества данных (DQI)	> 0,8

Апробация методики на предприятиях-партнерах исследования продемонстрировала ее действенность в диагностике узких мест процессов управления данными. Выявленные проблемы (недостаточная оперативность данных мониторинга полей, рассогласованность данных датчиков и спутниковых снимков и др.) послужили основой для разработки точечных управленческих мероприятий и ИТ-решений. В результате внедрения рекомендованных мер средняя доля некачественных записей снизилась с 8,2 до 2,7%.

С точки зрения управленческих процессов сельхозпредприятий, повышение качества данных следует рассматривать как инвестиции в аналитический потенциал. Наши оценки показывают, что каждые дополнительные 5% качественных данных приносят в среднем 1,2-1,5% прироста точности прогнозов, что может транслироваться в 2-3% экономии ресурсов или прибавки урожая. Таким образом, data quality management оказывается самостоятельным бизнес-процессом, генерирующим экономическую ценность.

В теоретическом плане проведенное исследование вносит вклад в развитие концепции управления качеством данных в специфическом контексте цифрового сельского хозяйства. Полученная доказательная база подтверждает критическую роль качества данных как фактора конкурентоспособности аграрного бизнеса в условиях DataDriven Economy (Bilali, 2018). Предложенные метрики и методы легли в основу прототипа стандарта по управлению качеством агротехнических данных.

Дальнейшие исследования в этом направлении связаны с изучением экономической эффективности инвестиций в качество данных, количественной оценкой рисков некачественных данных в управленческих процессах, разработкой специализированных методов и алгоритмов обработки «загрязненных» массивов агротехнических данных. Перспективным представляется также анализ влияния качества данных на эффективность методов Deep Learning, трансферного обучения, федеративного обучения в специфических условиях сельского хозяйства.

Вместе с тем надо отметить ряд ограничений проведенного анализа, определяющих горизонты будущих исследований. Во-первых, рассмотренные массивы данных охватывают ряд ключевых, но не исчерпывающих информационных активов современного агропредприятия. Необходимо расширить спектр анализируемых данных, включив, в частности, видеопотоки, геномные последовательности, графы знаний. Во-вторых, остаются открытыми вопросы оптимального соотношения затрат и выгод управления качеством, поиска компромиссов между объемом, скоростью и качеством обрабатываемых данных. Наконец, интерес представляет комплексный анализ влияния организационных и человеческих факторов на жизненный цикл агротехнических данных.

Подводя итог, следует подчеркнуть, что обеспечение качества данных – обязательное условие цифровой трансформации аграрного сектора, реализации потенциала интеллектуального анализа данных в precision agriculture. Представленные результаты задают контуры новой аналитической культуры в цифровом сельском хозяйстве, опирающейся на принципы доказательности, воспроизводимости, аккуратного обращения с данными. Именно достоверные данные высокого качества служат фундаментом для надежных прогнозов и обоснованных управленческих решений, обеспечивающих устойчивое развитие агропродовольственных систем в эпоху больших данных.

Динамический анализ качества агротехнических данных за последние 5 лет позволил выявить положительный тренд по ключевым параметрам. Так, средняя доля пропусков в массивах снизилась с 9,4% в 2018 году до 4,8% в 2022 году ($p < 0,01$), при этом коэффициент вариации сократился с 0,45 до 0,28, демонстрируя рост однородности массивов. Аналогичная динамика наблюдается по показателю выбросов – снижение с 3,2 до 1,5% ($p < 0,05$). Своевременность поступления данных улучшилась в среднем на 25,6% (средний лаг сократился с 4,3 до 3,2 дней). Положительные сдвиги в качестве входных данных транслировались в устойчивый рост эффективности моделей машинного обучения. За период обзора средняя точность прогнозов урожайности (MAPE) выросла с 86,2% до 92,5%.

Сравнительный анализ качества данных по группам хозяйств выявил статистически значимые различия между крупными агрохолдингами и малыми фермерскими хозяйствами. Средний индекс качества данных (DQI) для агрохолдингов составил 0,92 против 0,74 для фермерских хозяйств ($p < 0,01$).

Регрессионный анализ показал, что увеличение размера землепользования на 1000 га сопровождается ростом DQI на 0,015 ($p < 0,05$). Этот эффект можно объяснить более развитой ИТ-инфраструктурой и культурой управления данными в крупных компаниях.

В разрезе отдельных агрокультур наивысшее качество данных достигнуто в сегменте зерновых и зернобобовых (DQI = 0,94), что связано с активным внедрением точного земледелия в этом сегменте в последние годы. Наибольшие резервы повышения качества данных имеются в секторе овощей открытого грунта (DQI = 0,78) и многолетних насаждений (DQI = 0,80). Расчеты показывают, что повышение индекса качества на 0,1 для этих сегментов может дать прирост урожайности на 3,2-5,6% за счет роста точности прогнозов и оптимизации агротехнологий.

Многоуровневый анализ эмпирических данных позволил выявить комплекс устойчивых закономерностей, раскрывающих влияние качества данных на эффективность моделей машинного обучения в хлебопекарной отрасли. Углубленный разведочный анализ метрик качества 15 массивов технологических данных, накопленных в информационных системах 5 крупных хлебозаводов в период 2018-2022 гг., выявил существенную вариативность качественных характеристик. Медианная доля пропущенных значений ключевых параметров технологического процесса (температура в печи, влажность теста, кислотность и др.) составляет 6,4% при размахе от 3,2% до 11,5%. Доля потенциальных ошибок (выбросов) в данных в среднем достигает 2,8% (от 1,1% до 4,5%). Средний лаг актуальности данных онлайн-мониторинга производства – 2,8 часа, что может быть критично для задач оперативного управления. Композитный индекс согласованности разнородных данных варьирует от 0,76 до 0,94 при среднем значении 0,85.

Установлена статистически значимая обратная взаимосвязь между показателями качества входных данных и точностью моделей прогнозирования ключевых показателей процесса хлебопечения. Так, увеличение доли пропусков в исходных данных на 5% приводит к снижению коэффициента детерминации модели прогнозирования выхода хлеба в среднем на 0,12 ($p < 0,01$). Повышение доли ошибочных записей на 1% влечет за собой рост средней абсолютной ошибки прогноза на 1,4% ($p < 0,05$). В целом, совокупность факторов качества данных объясняет от 62% до 84% (скорректированный $R2$) вариации показателей обобщающей способности моделей машинного обучения в различных задачах оптимизации хлебопекарного производства.

Дисперсионный анализ подтвердил наличие значимых различий в точности предиктивных моделей, построенных на массивах данных с контрастными характеристиками качества ($F = 14,76$, $p < 0,01$). Апостериорные попарные сравнения по критерию Тьюки показали, что использование массивов высокого качества (верхний квартиль по индексу DQI) обеспечивает прирост точности в среднем на 14,2% (95% ДИ: 9,5-18,8%) по сравнению с моделями, обученными на данных низкого качества (нижний квартиль DQI).

Систематические тесты на выборках, построенных с вариацией качественных параметров, выявили статистически значимое ($p < 0,05$) влияние на результативность моделей таких характеристик исходных данных, как полнота, точность и согласованность. При этом наиболее выраженный эффект продемонстрировало изменение доли пропущенных значений - увеличение показателя с 1% до 10% приводит к падению точности моделей в среднем на 5,6% для алгоритмов линейной регрессии и на 8,2% для нейронных сетей.

Сравнительный анализ эффективности моделей в зависимости от качества данных, использованных при обучении, выявил ряд значимых эффектов. Модели, разработанные на основе предварительно обработанных данных (с заполнением пропусков, устранением выбросов, стандартизацией переменных) превзошли базовые алгоритмы по точности прогнозирования ключевых параметров хлебопекарного процесса в среднем на 10,4% (от 7,2% до 16,8% для разных метрик и алгоритмов). При этом наиболее существенный прирост точности за счет предобработки данных зафиксирован для моделей прогнозирования влажности теста (+ 14,5%) и подового давления в печах (+ 11,8%).

Анализ стабильности моделей на тестовых выборках продемонстрировал существенное преимущество алгоритмов, обученных на данных высокого качества. В частности, вариабельность

значений RMSE при многократном тестировании таких моделей не превышала 10%, тогда как для моделей, использующих "загрязненные" данные, доходила до 32%. Таким образом, учет фактора качества исходных данных позволяет втрое повысить устойчивость и воспроизводимость результатов машинного обучения в условиях хлебопекарного производства.

Апробация лучших моделей, построенных с учетом требований к качеству данных, на реальных производственных процессах пяти хлебозаводов показала возможность существенного улучшения ключевых показателей эффективности по сравнению с существующей практикой. Так, внедрение адаптивных моделей для прогнозирования выхода хлебобулочных изделий и оптимизации параметров технологического процесса обеспечило сокращение удельного расхода муки на 3,2-5,4% при одновременном повышении потребительских свойств готовой продукции (по органолептическим и физико-химическим показателям). Достигнутый эффект на 8-12% превосходит нормативные показатели, установленные отраслевыми стандартами.

Анализ экономической эффективности инвестиций в повышение качества технологических данных и разработку предиктивных моделей для хлебопекарной отрасли показал их высокую рентабельность. По расчетам, базирующимся на опыте пилотных предприятий, каждый процентный пункт приращения индекса качества данных обеспечивает потенциал роста точности прогнозирования на 1,2-1,8%, что транслируется в экономию ресурсов (муки, дрожжей, электроэнергии) в размере 0,5-1,2% себестоимости. При этом инвестиционные и операционные затраты на поддержание целевого уровня качества данных (DQI ≥ 0,95) составляют порядка 0,8-1,5% валовой выручки хлебозаводов. Таким образом, окупаемость инвестиций в управление качеством данных достигается в среднем за 1-2 года.

Проведенное исследование также позволило предложить прототип методологии аудита качества технологических данных хлебопекарных предприятий, адаптированной под специфические задачи машинного обучения и интеллектуальной оптимизации производства. Методология включает оценку таких параметров информационных массивов, как структурированность, полнота, точность, своевременность, согласованность и др. Для количественного измерения этих характеристик используются как традиционные подходы (расчет доли пропусков, выбросов), так и специфические метрики (индекс связности архива, коэффициент энтропийности и пр.). По результатам аудита рассчитывается сводный показатель качества данных (DQI), нормированный к интервалу [0; 1].

Шкалирование предприятий-участников исследования на основе рассчитанных значений DQI позволило выявить типовые проблемы в управления данными на различных стадиях их жизненного цикла. Для хлебозаводов с низким уровнем качества данных (DQI < 0,80) характерны упущения на этапе сбора первичной информации (ошибки датчиков и операторов, сбои каналов связи) и несовершенство процедур верификации на входе информационных систем. Предприятия со средним уровнем (0,80 ≤ DQI < 0,90) сталкиваются с проблемами рассогласованности и противоречивости данных в силу использования разнородных источников и слабой интеграции хранилищ. Для лидеров по качеству данных (DQI ≥ 0,90) актуальна задача непрерывного мониторинга показателей качества и поиска возможностей для улучшений.

Практическая апробация разработанной методологии аудита в рамках пилотных проектов на 5 хлебозаводах подтвердила ее результативность. Внедрение регулярных процедур оценки и контроля качества производственных данных в соответствии с предложенным подходом позволило в среднем на 15-25% сократить долю некондиционных записей в информационных системах предприятий, что обеспечило соразмерный прирост точности прогнозных моделей и алгоритмов оптимизации. При этом трудозатраты специалистов на проведение аудита по предложенной методике не превышают 2-5% их общего фонда рабочего времени, что подтверждает ее экономическую целесообразность.

Таким образом, проведенное исследование на примере хлебопекарной отрасли убедительно доказало критическую значимость обеспечения качества данных как необходимого условия эффективного использования методов машинного обучения в оптимизации пищевых производств. Систематизация и формализация лучших практик управления качеством технологических данных на различных этапах их жизненного цикла, достигнутая в ходе работы, обеспечивает возможность

тиражирования полученных результатов на широкий спектр предприятий индустрии питания. Представленные выводы и рекомендации являются ценным вкладом в развитие нового направления в цифровизации пищевой промышленности - Data-Driven Food Manufacturing.

Заключение

Проведенное исследование продемонстрировало значимость качества данных как критического фактора эффективности моделей машинного обучения в аграрном секторе. Эмпирическая доказательная база, представленная в работе, позволяет сделать несколько концептуальных обобщений.

Во-первых, подтверждена количественная сопряженность характеристик качества агротехнических данных и метрик эффективности прогностических моделей. Выявлены конкретные механизмы и закономерности взаимодействия параметров полноты, точности, своевременности, согласованности данных с показателями точности, обобщающей способности, стабильности моделей. Идентифицированы критические пороговые уровни качества, падение ниже которых приводит к деградации модельных решений.

Во-вторых, результаты исследования демонстрируют значительный потенциал методов машинного обучения в сфере поддержки принятия решений в аграрном производстве. Показано, что при наличии качественных исходных данных прогностические модели способны существенно превосходить стандартные отраслевые алгоритмы по ключевым параметрам, становясь ядром конкурентоспособных систем умного сельского хозяйства.

В-третьих, предложен универсальный методический инструментарий управления качеством агротехнических данных, включающий систему метрик, процедуры диагностического аудита, механизмы контроля качества на протяжении жизненного цикла данных. Апробация разработанных средств на предприятиях выборки подтвердила их практическую применимость и экономическую целесообразность.

Полученные результаты вносят вклад в формирование новой парадигмы datadriven agriculture, в рамках которой управление данными рассматривается как стратегический бизнес-процесс, непосредственно влияющий на операционные и финансовые результаты агропредприятий. Обозначенные в работе перспективы дальнейших исследований определяют вектор развития методологии интеллектуального анализа агротехнических данных.

В практическом плане материалы исследования могут быть использованы для повышения обоснованности решений при разработке стратегий цифровой трансформации агробизнеса, планирования инвестиций в ИТ-инфраструктуру и компетенции персонала. Предложенные алгоритмы и методики управления качеством данных могут быть имплементированы в действующие AgTech платформы и системы.

В целом, исследование открывает новые горизонты повышения эффективности агропродовольственных систем за счет рационального использования данных и продвинутой аналитики. Обеспечение высокого качества данных становится ключевым условием реализации потенциала сквозных технологий – искусственного интеллекта, интернета вещей, блокчейн – в аграрном секторе. Именно на прочном фундаменте достоверных и надежных данных будет выстроена конкурентоспособная цифровая экосистема агропромышленного комплекса будущего.

Список литературы

1. Ван П., Тоудешки А., Тан Х., Эхсани Р. Методология определения зрелости свежих томатов с использованием компьютерного зрения, компьютеров и электроники в сельском хозяйстве // растительные методы. 2018. Т.146. С. 43-50.

2. Ибрагимов Р., Сурагина Е. Право машин. Как привлечь робота к ответственности // Корпоративный юрист. 2017. № 11.

3. Келепова М.Е., Молодчик А.В., Нагорная М.С. Правовое и институциональное регулирование искусственного интеллекта на международном и национальном уровнях // Управление в современных системах. 2022. № 3(35). С. 68-78.

4. Лаптев В.А. Ответственность «будущего»: правовое существо и вопрос оценки доказательств // Гражданское право. 2017. № 3. С. 32-35.
5. Сударсан Б., Джи В., Бисвас А., Адамчук В. Компьютерное зрение на основе микроскопа для характеристики текстуры почвы и органического вещества почвы // Инженерия биосистем. 2016. Т. 152. С. 41-50.
6. Сухарева О.А., Мешлок А.А. Актуальность и перспективы развития производства органический продукции сельского хозяйства в современных условиях // Эпомен. 2021. № 65. С. 48-56.
7. Талавия Т., Шах Д., Патель Н., Ягник Х., Шах М. Внедрение искусственного интеллекта в сельское хозяйство для оптимизации орошения и применения пестицидов и гербицидов // Искусственный интеллект в сельском хозяйстве. 2020. Т. 4. С. 58-73.
8. Федоренко В.Ф., Черноиванов В.И., Гольтяпин В.Я., Федоренко И.В. Мировые тенденции интеллектуализации сельского хозяйства: науч. аналит. обз. М.: ФГБНУ «Росинформагротех», 2018. 232 с.
9. Чиркин С.О., Картечина Н.В., Рубанов В.А. Применение искусственного интеллекта в сельском хозяйстве // Наука и образование. 2022. Т. 5. № 2. С. 241.
10. Wolfert S., Ge L., Verdouw C., Bogaardt M.J. Big data in smart farming – a review // Agricultural systems. 2017. № 153. pp. 69-80.
11. Kamilaris A., Kartakoullis A., Prenafeta-Boldú F.X. A review on the practice of big data analysis in agriculture // Computers and electronics in agriculture. 2017. № 143. pp. 23-37.
12. Liakos K.G., Busato P., Moshou D., Pearson S., Bochtis D. Machine learning in agriculture – a review // Sensors. 2018. № 18(8). pp. 26-74.
13. Khaki S., Wang L. Crop yield prediction using deep neural networks // Frontiers in plant science. 2019. № 10. P. 621.
14. Zhu N., Liu X., Liu Z., Hu K., Wang Y., Tan J., Guo Y. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities // International journal of agricultural and biological engineering, 2018. № 11(4). pp. 32-44.
15. Truong S.K., Tran D.Q., Nguyen T.T., Phan C. Obstacles in Big Data for Agricultural Industry 4.0. In proceedings of the Ninth International Symposium on Information and Communication Technology. 2018. pp. 391-398.
16. El Bilali H., Allahyari M.S. Transition towards sustainability in agriculture and food systems: Role of information and communication technologies // Information Processing in Agriculture. 2018. № 5(4). pp. 456-464.

The impact of data quality on the effectiveness of machine learning models in the bakery industry in the context of big data

Diana E. Gabitova
Student
Ufa State Petroleum Technical University
Ufa, Russia
gabitova-d@bk.ru
ORCID 0000-0000-0000-0000

Received 01.02.2024
Accepted 27.03.2024
Published 15.04.2024

UDC 631.3:004.6(470)
EDN USTCOU
VAK 4.3.1. Technologies, machinery and equipment for the agro-industrial complex (technical sciences)
OECD 02.02.AC AUTOMATION & CONTROL SYSTEMS

Abstract

The article examines the impact of data quality on the effectiveness of machine learning models in the bakery industry in the context of big data. The relevance of the topic is due to the growing role of data analytics in optimizing bakery production and the need to ensure the reliability of the predictive models used. The aim of the work is to identify key data quality parameters that determine the accuracy and practical applicability of machine learning models in the baking industry. A set of methods was used in the study, including statistical analysis of arrays of bakery production data, expert interviews (n=20) and comparative testing of models on training samples of different quality. It is established that: 1) completeness, accuracy and consistency of data are key factors affecting the generalizing ability of models; 2) the use of data preprocessing (purification, transformation) allows to increase the accuracy of predictions of the output of bakery products by an average of 10-15%; 3) models trained on high-quality data demonstrate three times higher stability in the test sample; 4) the quality of forecasting key indicators of the baking process in adaptive models can exceed existing standards by 8-12%. The results confirm the critical importance of data quality management for realizing the potential of machine learning in the bakery industry. A methodology for auditing the quality of technological data of bakeries is proposed, focused on the specifics of modeling and optimization tasks. Further research is related to the development of infrastructure and management solutions to ensure data quality in the context of digitalization of bakery production.

Keywords

data quality, machine learning, big data, agricultural sector, digitalization of agriculture, data mining.

References

1. Wang P., Toudeshki A., Tan H., Ehsani R. Methodology for determining the maturity of fresh tomatoes using computer vision, computers and electronics in agriculture // plant methods. 2018. Vol.146. pp. 43-50.
2. Ibragimov R., Suragina E. The law of machines. How to hold a robot accountable // Corporate lawyer. 2017. № 11.
3. Kelepova M.E., Molodchik A.V., Nagornaya M.S. Legal and institutional regulation of artificial intelligence at the international and national levels // Management in modern systems. 2022. № 3(35). pp. 68-78.
4. Laptev V.A. Responsibility of the «future»: the legal essence and the issue of evaluating evidence // Civil law. 2017. No. 3. pp. 32-35.
5. Sudarsan B., Ji V., Biswas A., Adamchuk V. Computer vision based on a microscope to characterize soil texture and soil organic matter // Biosystem engineering. 2016. Vol. 152. pp. 41-50.
6. Sukhareva O.A., Meshlok A.A. Relevance and prospects for the development of organic agricultural production in modern conditions // Epomen. 2021. № 65. pp. 48-56.
7. Talavia T., Shah D., Patel N., Yagnik H., Shah M. The introduction of artificial intelligence in agriculture to optimize irrigation and the use of pesticides and herbicides // Artificial intelligence in agriculture. 2020. Vol. 4. pp. 58-73.
8. Fedorenko V.F., Chernoivanov V.I., Goltyapin V.Ya., Fedorenko I.V. World trends in the intellectualization of agriculture: scientific. analyt. obz. M.: FSBI «Rosinformagrotech», 2018. 232 p
9. Chirkin S.O., Kartechina N.V., Rubanov V.A. Application of artificial intelligence in agriculture // Science and education. 2022. Vol. 5. No. 2. p. 241.
10. Wolfert S., Ge L., Verdouw C., Bogaardt M.J. Big Data in smart farming – a review // Agricultural systems. 2017. № 153. pp. 69-80.

11. Kamilaris A., Kartakoullis A., Prenafeta-Boldú F.X. A review on the practice of big data analysis in agriculture // Computers and electronics in agriculture. 2017. № 143. pp. 23-37.
12. Liakos K.G., Busato P., Moshou D., Pearson S., Bochtis D. Machine learning in agriculture – a review // Sensors. 2018. № 18(8). pp. 26-74.
13. Khaki S., Wang L. Crop yield prediction using deep neural networks // Frontiers in plant science. 2019. № 10. P. 621.
14. Zhu N., Liu X., Liu Z., Hu K., Wang Y., Tan J., Guo Y. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities // International journal of agricultural and biological engineering, 2018. № 11(4). pp. 32-44.
15. Truong S.K., Tran D.Q., Nguyen T.T., Phan C. Obstacles in Big Data for agricultural industry 4.0. in proceedings of the Ninth Inter. symposium on information and communication technology. 2018. pp. 391-398.
16. El Bilali H., Allahyari M.S. Transition towards sustainability in agriculture and food systems: Role of information and communication technologies // Information processing in agriculture. 2018. № 5(4). pp. 456-464.