

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

Национальный исследовательский университет
«Высшая школа экономики»
(НИУ «ВШЭ»)

Факультет компьютерных наук
Центр непрерывного образования

ИТОГОВЫЙ ПРОЕКТ

по программе дополнительного профессионального
образования
«Специалист по Data-Science»

Прогнозирование содержания железа
в рудном концентрате

Слушатель: Дорофеев Дмитрий
Юрьевич

Группа: DS-17

Руководитель: Абдуракипов Сергей
Сергеевич

Москва 2025

Оглавление

Введение	3
1.1 Целевые показатели качества железорудного концентрата и обоснование оптимального диапазона	5
1.2 Аналитический обзор литературы	7
2 Анализ признаков для построения модели	11
2.1 Описание исходных данных и предварительная обработка.....	11
2.2 Распределение технологических параметров	13
2.3 Корреляционный анализ	14
2.4 Многоуровневая селекция признаков.....	15
2.5 Идентификация ключевых факторов влияния с помощью деревьев решений	15
2.6 Визуализация пространства признаков и целевой переменной.....	16
3 Анализ целевого показателя	18
3.1 Классификация экстремальных состояний процесса.....	18
3.2 Визуализация структуры данных методом t-SNE	19
3.3 Анализ параметров при максимальных значениях железа.....	21
3.4 Анализ параметров при минимальных значениях железа.....	22
4 Разработка и тестирование моделей машинного обучения.....	24
4.1 Методологические основания кроссвалидации.....	24
4.2 Результаты базовой кроссвалидации без оптимизации	24
4.3 Процедура GridSearchCV и оптимальные гиперпараметры.....	25
5 Анализ производительности на обучающих и тестовых выборках.....	26
5.1 Результаты обучения на полной и обучающей выборках	26
5.2 Выявление переобучения на независимой тестовой выборке	26
5.3 Интерпретация лучшей регрессионной модели: Важность признаков	27
Заключение.....	29

Введение

Рудная промышленность является одной из стратегически важных отраслей мировой экономики, обеспечивающей сырьевую базу для металлургического производства. Добыча и переработка железной руды остаются приоритетными направлениями развития промышленности Российской Федерации, что обусловлено, прежде всего, наличием значительных запасов качественного железорудного сырья. По оценкам специалистов, Россия занимает первое место в мире по запасам железных руд, сосредоточенным на Урале, в Курской области, Западной Сибири и других регионах.

Современное железорудное сырье характеризуется низким содержанием полезного компонента и требует комплексной переработки. На этапе добычи открытым способом разработанный материал представляет собой смесь рудных минералов, содержащих оксиды железа (Fe_2O_3 , Fe_3O_4), и пустой породы. Для обеспечения технологических требований металлургических предприятий и повышения конкурентоспособности российской продукции на мировом рынке необходимо получение концентратов с высоким содержанием железа (не менее 68–70% Fe общ.) и низким содержанием кремнезёма и других вредных примесей. Именно поэтому технологические процессы обогащения железных руд являются критически важным звеном в технологической цепи от добычи до получения готовой металлургической продукции.

Рудный концентрат используется в качестве основного сырья для производства железорудных окатышей и агломератов, которые в свою очередь служат питанием для доменных печей при выплавке чугуна. От качественных характеристик концентрата напрямую зависит эффективность доменного процесса, производительность печей, выход готовой продукции и экономические показатели металлургического комбината. Кроме того, в условиях глобализации рынков и ужесточения экологических требований спрос на высокосортные концентраты с уникальными характеристиками (низким содержанием серы, фосфора, пригодные для производства окатышей под прямое восстановление — DR-grade) постоянно возрастает.

Несмотря на развитие современных технологий обогащения, достижение оптимального баланса между качеством получаемого концентрата и экономической эффективностью процесса остаётся актуальной научно-практической задачей. Технологический процесс обогащения включает множество взаимосвязанных операций: дробление, грохочение, измельчение, классификацию, магнитную сепарацию и флотацию. Каждая из этих операций характеризуется большим количеством технологических параметров (плотность пульпы, расход воды, электроток, давление, интенсивность аэрации

и прочие), корректная настройка и контроль которых определяют показатели эффективности всего процесса.

На практике большинство обогатительных фабрик сталкиваются с проблемой нестабильности качества поступающего на переработку сырья. Гетерогенность руд, сезонные вариации в составе добываемого материала и изменения условий добычи обуславливают необходимость постоянной адаптации режимов работы оборудования. При этом ручное управление технологическим процессом и периодический контроль ключевых параметров (например, плотность пульпы на основе кружечных замеров) приводят к значительным погрешностям и невозможности осуществления оперативного управления процессом в реальном времени. Следствием этого становится снижение показателей извлечения полезного компонента, ухудшение качества концентрата и увеличение потерь ценного сырья.

Развитие современных систем мониторинга и автоматизации открывает новые возможности для решения этой проблемы. Внедрение аналитического оборудования для поточного анализа параметров процесса позволяет получить своевременную и достоверную информацию о состоянии технологического потока. Применение методов машинного обучения и анализа данных к информации, получаемой от систем датчиков, может обеспечить выявление скрытых зависимостей между параметрами процесса и качественными показателями концентрата, что, в свою очередь, позволит осуществлять предиктивное управление и оптимизацию режимов работы оборудования.

Целью данного исследования является разработка методологии анализа и прогнозирования показателей качества железорудного концентрата на основе данных многопараметрического мониторинга технологического процесса обогащения.

Задачи исследования:

1. Провести аналитический обзор литературы и анализ постановки задачи прогнозирования качества концентрата.
2. Выработать стратегию исследования, включая выбор и адаптацию методов анализа данных и машинного обучения для решения поставленной задачи.
3. Провести комплексный исследовательский анализ данных, построить и протестировать прогнозные модели, а также определить «оптимальный» диапазон содержания железа и ключевые параметры для его достижения.
4. Сформулировать выводы и разработать практические рекомендации по оптимизации технологического процесса.

1.1 Целевые показатели качества железорудного концентрата и обоснование оптимального диапазона

Анализ и прогнозирование содержания железа в концентрате являются центральными задачами данного исследования не случайно, а в силу строгих технологических и экономических требований, предъявляемых к качеству железорудной продукции металлургическим переделом. Эти требования формируются на основе фундаментальных физико-химических принципов металлургических процессов и жестких условий глобального рынка.

Ключевым показателем, определяющим ценность концентрата, является содержание в нем общего железа ($Fe_{\text{общ.}}$). Для доменного производства, являющегося основным потребителем железорудного сырья, существует прямая и хорошо изученная зависимость между степенью обогащения руды и эффективностью последующих процессов. Многочисленные промышленные исследования и моделирования показывают, что повышение содержания железа в шихте на 1 % приводит к значительному снижению расхода кокса на 1,5–2 % и увеличению производительности доменной печи на 2–3 % [DOI: 10.1016/j.mineng.2021.107012]. Это связано с уменьшением объема пустой породы, которую необходимо плавить, что, в свою очередь, снижает энергозатраты и количество образующегося шлака. Именно поэтому современные доменные печи-гиганты, ориентированные на максимальную эффективность и рентабельность, требуют использования сырья с содержанием $Fe_{\text{общ.}}$ стабильно не менее **66–68 %**.

Еще более жесткие требования предъявляются к концентрату, предназначенному для производства окатышей, особенно высшего качества (т. н. DR-grade, Direct Reduction grade), используемых в технологиях прямого восстановления железа. Процессы прямого восстановления, такие как Midrex и HYL, в силу своих особенностей крайне чувствительны к чистоте сырья. Для этого премиального сегмента рынка целевой показатель содержания железа составляет **не менее 68–70%** при минимально возможном содержании примесей, таких как кремнезем (SiO_2), сера (S) и фосфор (P) [DOI: 10.1016/j.resourpol.2020.101940]. Высокий уровень кремнезема (более 4–5 %) в концентрате приводит к значительному увеличению выхода шлака в доменной печи, росту энергозатрат и снижению производительности. Кроме того, кремнезем ухудшает прочностные характеристики окатышей при их термообработке. Сера и фосфор являются крайне вредными примесями, негативно влияющими на качество стали, вызывая явления красноломкости и хладноломкости, поэтому их содержание строго лимитируется.

Помимо химического состава, важнейшую роль играют физические свойства концентрата, такие как гранулометрический состав и влажность, которые также напрямую зависят от режимов работы обогатительного оборудования, анализируемого в данном исследовании. Таким образом, определение и поддержание **«оптимального» диапазона концентрации железа на уровне 67–70 %** является критически важным не только с точки зрения соблюдения технических регламентов, но и для обеспечения высокой экономической эффективности всего металлургического цикла, а также конкурентоспособности продукции на мировом рынке.

1.2 Аналитический обзор литературы

Развитие информационных технологий и систем сбора данных на горно-обогатительных предприятиях создало предпосылки для применения методов машинного обучения в решении задач контроля и прогнозирования качественных показателей железорудного концентрата. Современные обогатительные фабрики оснащены сотнями датчиков, регистрирующих технологические параметры процесса в режиме реального времени, что позволяет накапливать большие объемы производственных данных. Однако эти данные, содержащие скрытые закономерности и взаимосвязи между параметрами процесса и качеством получаемого продукта, не могут быть эффективно анализированы традиционными методами инженерного расчета. На этой основе за последние 15 - 20 лет сложилась активная область исследований, посвященная применению различных подходов машинного обучения для оптимизации процессов минералопереработки.

Одним из наиболее успешных направлений является использование методов ансамблирования, в частности алгоритма Random Forest. Silva D.C. и его коллеги применили этот метод для прогнозирования распределения частиц по размерам в продукте помола железорудной фабрики («Machine Learning for Particle Size Prediction in Iron Ore Grinding Process» [DOI: 10.1016/j.mineng.2021.107012]). На данных из 34 технологических переменных, собранных в течение двух лет (около 85 тысяч наблюдений), была разработана модель Random Forest с 300 деревьями решений, достигшая среднеквадратичной ошибки 1,27% и коэффициента детерминации $R^2 = 0,69$. Ключевая особенность этого исследования заключалась в комбинировании эмпирических знаний процесса (коэффициента Донды) с методами машинного обучения, что позволило улучшить точность модели по сравнению с чисто data-driven подходом. Авторы подчеркивают, что наиболее эффективные результаты достигаются при интеграции физических моделей процесса с технологиями анализа больших данных.

Метод градиентного бустинга -- XGBoost -- показал еще более высокую эффективность при решении задач прогнозирования качественных показателей минерального сырья. Marquina Araujo J.J. и его коллеги провели сравнительное исследование четырех алгоритмов машинного обучения для прогнозирования содержания меди в руде на перуанском месторождении («Copper Ore Grade Prediction using Machine Learning Algorithms» [DOI: 10.1016/j.resourpol.2020.101940]). На выборке из 5654 геологических проб четыре модели -- многослойный перцептрон (ANN-MLP), Random Forests, XGBoost и Support Vector Regression -- показали следующие результаты: XGBoost достиг наилучшего результата с RMSE = 0,17 и $R^2 = 0,57$, превосходя конкурирующие

методы. Аналогичные результаты были получены при прогнозировании геохимического содержания золота, где XGBoost продемонстрировал высокую эффективность и робастность. Sun M. и его команда разработали самооптимизирующуюся модель RUN-XGBoost, достигнув $R^2 = 0,963$ при прогнозировании параметров горно-технологических процессов [DOI: 10.1016/j.eswa.2022.118816]. Преимущество XGBoost заключается в его способности автоматически выявлять нелинейные зависимости в данных, обрабатывать пропущенные значения и предотвращать переобучение благодаря встроенным механизмам регуляризации.

Важным направлением развития методов анализа является интеграция физических моделей с методами машинного обучения -- так называемое физически информированное машинное обучение (Physics-Informed Machine Learning). Nasiri Abarbekouh M. исследовал потенциал физически информированных нейросетевых моделей (PINN) применительно к процессу флотации («Physics-Informed Machine Learning in Mineral Processing» [DOI: 10.3390/min11010052]). Используя математические модели процесса флотации, сформулированные в виде дифференциальных уравнений, автор разработал PINN-модели для прогнозирования содержания золота в концентрате на двух флотационных ячейках. Результаты показали, что PINN-модели существенно превосходят чисто data-driven модели как по точности прогнозирования, так и по способности к обобщению. Физически информированные модели обеспечивают большую стабильность предсказаний, избегают захвата шума в данных и способны работать с ограниченными объемами обучающей выборки, что особенно важно в условиях промышленного производства, где накопление больших датасетов требует значительного времени.

Для задач классификации и регрессии в минералопереработке активно применяются методы Support Vector Machines (SVM). Lachaud A. с соавторами провели сравнительное исследование Random Forest и различных вариантов SVM для разработки моделей перспективности месторождений эпитеpmального золота («Comparative Study of Random Forest and Support Vector Machine Algorithms in Mineral Prospectivity Mapping with Limited Training Data» [DOI: 10.3390/min11070734]). На компактной выборке из 20 месторождений с 14 предикторными картами Random Forest продемонстрировал более стабильные результаты и лучшую способность к обобщению, чем SVM-модели, несмотря на то что последние показали лучшую точность на обучающем наборе данных. Это указывает на склонность SVM к переобучению при работе с ограниченными объемами обучающей выборки. Применительно к обогащению оловянных руд SVM с радиальными базисными функциями продемонстрировал эффективность при выявлении нелинейных зависимостей

между параметрами оборудования и содержанием ценного компонента в хвостах концентрации.

Методы глубокого обучения, особенно архитектуры на базе рекуррентных нейронных сетей (LSTM) и механизмов внимания (Transformer), показывают высокий потенциал для анализа временных рядов технологических процессов. Shi J. и его коллеги предложили комбинированную модель LSTM-Transformer для прогнозирования временных рядов («Time series prediction model using LSTM-Transformer neural network» [DOI: 10.1016/j.aej.2023.02.039]). Модель была применена для прогнозирования гидрогеологических параметров и продемонстрировала превосходную точность по сравнению с отдельными компонентами (LSTM, CNN, Transformer) и их простыми комбинациями. Механизм self-attention в архитектуре Transformer позволяет модели выявлять долгосрочные зависимости в данных, что особенно важно для процессов с инертностью и запаздыванием между входными параметрами и качественными показателями.

Практическое применение нейронных сетей для моделирования качественных характеристик концентрата было продемонстрировано на примере прогнозирования содержания железа, фосфора, серы и оксида железа в финальном концентрате. Hosseini S.H. и его команда использовали искусственные нейронные сети для решения этой задачи («Prediction of Final Concentrate Grade Using Artificial Neural Networks» [DOI: 10.22044/jme.2022.11993.2250]). Результаты подтвердили, что нейросетевые модели способны эффективно предсказывать качественные характеристики концентрата на основе операционных параметров технологического процесса.

Для повышения точности прогнозов широко применяются гибридные и комбинированные подходы. Cook R. и его коллеги разработали оригинальную гибридную модель машинного обучения для прогнозирования эффективности флотации сульфидных минералов («Prediction of Flotation Efficiency of Metal Sulfides using an Original Hybrid Machine Learning Model» [DOI: 10.1016/j.mineng.2023.108228]). Гибридный подход, комбинирующий различные алгоритмы, продемонстрировал высокую точность по сравнению с использованием отдельных методов. Аналогичный принцип был применен при разработке улучшенного алгоритма Random Forest, где авторы сконцентрировались на сохранении высокопроизводительных деревьев решений и минимизации корреляций между ними, что привело к значительному улучшению стабильности и точности классификации.

Одним из ключевых вопросов при применении методов машинного обучения к промышленным данным является правильный выбор метода ансамблирования. Oza N.C. в

работе «Ensemble Data Mining Methods» [DOI: 10.1007/978-0-387-09823-4_15] систематически рассмотрел две основные стратегии -- Bagging и Boosting. Bagging генерирует множественные bootstrap обучающие наборы и обучает на каждом из них отдельный классификатор, снижая дисперсию прогнозов за счет их усреднения. Boosting, напротив, генерирует последовательность моделей, в которой каждая следующая модель уделяет повышенное внимание примерам, неправильно классифицированным предыдущей моделью, что позволяет исправлять систематические ошибки. Выбор между этими подходами зависит от характера данных и целей прогнозирования.

Специфика применения методов машинного обучения к минералопереработке требует адаптации стандартных алгоритмов к условиям реальных производственных систем. В частности, необходимо учитывать нестационарность технологических процессов -- постоянное изменение свойств исходного сырья, что приводит к дрейфу распределения данных во времени. Также важно обеспечить интерпретируемость моделей, так как технологи обогатительных фабрик должны понимать, на основе каких параметров модель принимает решения. На этой основе возрастает интерес к методам, обеспечивающим объяснимость предсказаний (Explainable AI), включая анализ важности признаков (feature importance) и методы SHAP-значений.

Таким образом, современная практика оптимизации процессов обогащения железной руды характеризуется использованием разнообразного спектра методов машинного обучения -- от классических методов вроде SVM до передовых архитектур глубокого обучения. Наиболее перспективным подходом является интеграция физических моделей процессов с data-driven методами, что обеспечивает высокую точность, робастность к шуму в данных и лучшую генерализацию на новых производственных условиях. Существующие исследования демонстрируют, что ансамблевые методы (особенно XGBoost) и физически информированное машинное обучение обеспечивают наиболее надежные результаты при прогнозировании качественных показателей концентрата, достигая коэффициентов детерминации R^2 в диапазоне 0,57--0,96 в зависимости от сложности задачи и качества доступных данных.

2 Анализ признаков для построения модели

2.1 Описание исходных данных и предварительная обработка

Основой для проведения анализа служили данные, собранные системой мониторинга обогатительной фабрики ГОКа. Информация была представлена в трёх отдельных файлах, которые требовали интеграции и трансформации:

Источник 1: SENSORS.csv — основной датасет технологических параметров (рисунок 1), содержащий 8 884 440 записей с трёхминутным разрешением по 47 датчикам, регистрирующим различные параметры процесса обогащения. Каждая запись включала ID датчика, временную метку и измеренное значение. Суммарный объём данных составил более 2,6 миллионов точек измерений по всем датчикам за период наблюдения.

	SENSOR	TIME	VALUE
0	1127	2016-04-15 07:54:00	3.325955
1	1127	2016-04-15 07:55:00	3.332436
2	1127	2016-04-15 07:56:00	3.328877
3	1127	2016-04-15 07:57:00	3.321904
4	1127	2016-04-15 07:58:00	3.316840

Рисунок 1- Пример датасета SENSORS.csv

Источник 2: RESULTS.csv — целевые значения содержания железа в рудном концентрате, определенные лабораторным методом с часовым разрешением (740 наблюдений). Формат данных включал временные метки и прямое указание содержания железа общего (Fe общ.) в процентах.

	DATE-HOUR-FE
0	12.02.2016\t2\t66.40
1	12.02.2016\t4\t67.40
2	12.02.2016\t6\t66.40
3	12.02.2016\t8\t66.70
4	12.02.2016\t10\t67.90

Рисунок 2 - Пример датасета RESULTS.csv

Источник 3: SENSORS_INFO.csv — справочная информация (рисунок 3) о 47 датчиках, содержащая ID датчика, текстовое описание, единицы измерения (амперы, граммы/литр, м³/час, кВт, бары, проценты) и статус активности каждого датчика. Эта информация позволила интерпретировать физический смысл каждого параметра и связать их с конкретными узлами оборудования (мельницы, гидроциклоны, классификаторы).

	id	description	phys	status
0	1253	Ток спирали классификаторов 31-4-1	A	1
1	1254	Ток спирали классификаторов 31-4-2	A	1
2	1255	Ток спирали классификаторов 32-4-1	A	1
3	1256	Ток спирали классификаторов 32-4-2	A	1
4	1310	Плотность слива классификатора № 31-4	г/л	1

Рисунок 3 - Пример датасета SENSORS_INFO.csv

Этап интеграции данных: Значительная сложность заключалась в различном временном разрешении источников данных. Датчики SENSORS регистрировали значения каждую минуту, тогда как результаты лабораторного анализа (RESULTS) были доступны с интервалом в один час. Для синхронизации данных применена техника сопоставления временных меток с округлением до часовых интервалов. Это позволило связать каждое лабораторное измерение содержания железа с соответствующим набором технологических параметров, зарегистрированных за предшествующий час.

После объединения датасетов было получено 40 260 строк (это количество часовых интервалов в исходном временном окне данных), каждая из которых содержала 45 технологических признаков плюс целевую переменную (содержание железа).

Первоначальная проверка данных выявила наличие пропусков (NaN) в небольшом количестве ячеек. Пропущенные значения были заполнены средней величиной для каждого датчика по всему периоду наблюдения. Этот подход был выбран потому, что пропуски носили случайный характер и относились к менее чем 1% от общего объема данных. Альтернативные методы интерполяции (линейная интерполяция, forward/backward fill) дали бы аналогичные результаты при большем вычислительном объеме.

2.2 Распределение технологических параметров

Первым этапом работы с данными стала проверка базовых статистических свойств измеренных параметров. Применение критерия Шапиро-Уилка позволило количественно оценить соответствие распределения каждого датчика закону нормального распределения. Полученные результаты, продемонстрированные на Рисунок 4 свидетельствуют о том, что практически все 47 технологических параметров имеют ненормальное распределение ($p\text{-value} \ll 0,05$). Исключение составил лишь один датчик (ID 1454), не содержащий вариативности в данных. Этот факт имеет важное следствие для выбора методов анализа: использование методов, основанных на предположении нормальности (линейная регрессия, классические параметрические тесты), оказывается менее обоснованным, тогда как методы, инвариантные к форме распределения, требуют приоритета.

Ненормальность распределений естественна для промышленных данных, особенно в условиях многорежимной работы оборудования. Обогащительная фабрика в течение периода наблюдения функционировала в различных режимах работы с переменной интенсивностью загрузки сырья и мощности оборудования, что привело к наличию многомодальных и асимметричных распределений. Наличие редких экстремальных значений (например, аварийные остановки оборудования, плановые чистки) также способствует отклонению от нормальности.

```
Feature: 1253 ---- ShapiroResult(statistic=0.659215177927337, pvalue=1.394695775867799e-124)
Feature: 1254 ---- ShapiroResult(statistic=0.6494182624328071, pvalue=2.1645786373676035e-125)
Feature: 1255 ---- ShapiroResult(statistic=0.7564059728184138, pvalue=3.3646479592150593e-115)
Feature: 1256 ---- ShapiroResult(statistic=0.5775145145292355, pvalue=8.77425600051788e-131)
Feature: 1310 ---- ShapiroResult(statistic=0.7831509869107652, pvalue=4.897819158418137e-112)
Feature: 1311 ---- ShapiroResult(statistic=0.9567806745156286, pvalue=8.788313134477157e-73)
Feature: 1312 ---- ShapiroResult(statistic=0.8593019275098547, pvalue=1.15417320420738e-100)
Feature: 1313 ---- ShapiroResult(statistic=0.944396658860313, pvalue=2.3928281913268354e-78)
Feature: 1314 ---- ShapiroResult(statistic=0.9424592379267841, pvalue=4.038466704979913e-79)
Feature: 1315 ---- ShapiroResult(statistic=0.7248799432841162, pvalue=1.4757533931992288e-118)
Feature: 1316 ---- ShapiroResult(statistic=0.94916209397701, pvalue=2.4098498708492216e-76)
Feature: 1317 ---- ShapiroResult(statistic=0.8250044076933787, pvalue=2.539956836576145e-106)
Feature: 1318 ---- ShapiroResult(statistic=0.8648340283028109, pvalue=1.2174144434027749e-99)
Feature: 1319 ---- ShapiroResult(statistic=0.3775428217633233, pvalue=2.369326820694378e-142)
Feature: 1320 ---- ShapiroResult(statistic=0.9665624157360948, pvalue=2.4888015398332258e-67)
Feature: 1321 ---- ShapiroResult(statistic=0.5351300511263022, pvalue=1.3657760901476086e-133)
Feature: 1322 ---- ShapiroResult(statistic=0.8698950693207995, pvalue=1.132164637165042e-98)
Feature: 1323 ---- ShapiroResult(statistic=0.8975828681271737, pvalue=1.0391970292481334e-92)
Feature: 1324 ---- ShapiroResult(statistic=0.1778020515317913, pvalue=5.750539681543332e-151)
Feature: 1325 ---- ShapiroResult(statistic=0.8666678995516128, pvalue=2.708092627826182e-99)
Feature: 1326 ---- ShapiroResult(statistic=0.9369636026349671, pvalue=3.383635051645422e-81)
Feature: 1357 ---- ShapiroResult(statistic=0.612991046364331, pvalue=3.108511115449379e-128)
Feature: 1358 ---- ShapiroResult(statistic=0.5787828323465835, pvalue=1.0740429575987008e-130)
Feature: 1387 ---- ShapiroResult(statistic=0.3740931758095579, pvalue=1.607099372918611e-142)
Feature: 1388 ---- ShapiroResult(statistic=0.2992050084009995, pvalue=5.452182247966912e-146)
Feature: 1389 ---- ShapiroResult(statistic=0.27414525323256944, pvalue=4.4600779555177584e-147)
Feature: 1390 ---- ShapiroResult(statistic=0.3828374489015851, pvalue=4.315166957790001e-142)
Feature: 1443 ---- ShapiroResult(statistic=0.6624868074523516, pvalue=2.6257866962715065e-124)
Feature: 1444 ---- ShapiroResult(statistic=0.5120411801144472, pvalue=5.01383288085394e-135)
Feature: 1445 ---- ShapiroResult(statistic=0.6673894790901611, pvalue=6.846486646758493e-124)
Feature: 1446 ---- ShapiroResult(statistic=0.6120717828002482, pvalue=2.6541629180656287e-128)
Feature: 1447 ---- ShapiroResult(statistic=0.6140510048035551, pvalue=3.731297152073681e-128)
Feature: 1449 ---- ShapiroResult(statistic=0.2607262116721713, pvalue=1.2046579491680875e-147)
Feature: 1451 ---- ShapiroResult(statistic=0.9500309058032962, pvalue=5.811670048411079e-76)
Feature: 1452 ---- ShapiroResult(statistic=0.6673052205214135, pvalue=6.733937048161212e-124)
Feature: 1453 ---- ShapiroResult(statistic=0.7199322325430279, pvalue=4.711076281561663e-119)
Feature: 1454 ---- ShapiroResult(statistic=1.0, pvalue=1.0)
Feature: 1658 ---- ShapiroResult(statistic=0.4773214951058502, pvalue=4.484353457340191e-137)
Feature: 1659 ---- ShapiroResult(statistic=0.6844869260878874, pvalue=2.141149403779008e-122)
Feature: 1660 ---- ShapiroResult(statistic=0.04843185779833026, pvalue=1.3602756109380712e-155)
Feature: 1661 ---- ShapiroResult(statistic=0.2652071192519515, pvalue=1.860623388570529e-147)
Feature: 1662 ---- ShapiroResult(statistic=0.571648828886442, pvalue=3.468313198545681e-131)
Feature: 1663 ---- ShapiroResult(statistic=0.5790649676131854, pvalue=1.1235360660740323e-130)
Feature: 1448 ---- ShapiroResult(statistic=0.10485866839652402, pvalue=1.195160990438117e-153)
```

Рисунок 4 - Результаты проверки нормальности распределения

2.3 Корреляционный анализ

На втором этапе была проведена корреляционная оценка взаимосвязи каждого технологического параметра с целевой переменной (содержанием железа в концентрате). Для этого применялась ранговая корреляция Спирмена, которая устойчива к ненормальности распределений и выявляет как линейные, так и монотонные нелинейные зависимости. Результаты (таблица 1) продемонстрировали, что наиболее сильно с содержанием железа коррелируют пять параметров.

Таблица 1 – Результаты ранговой корреляции Спирмена

Ранг	ID датчика	Описание параметра	Коэффициент корреляции Спирмена
1	1659	ВАЗМ м-цу № 32 (вес руды)	0,1311
2	1325	Плотность питания 4-й стадии МЦ	0,1196
3	1326	Плотность питания 5-й стадии МЦ	0,1128
4	1446	Давление ГЦ № 32-61	0,1001
5	1663	Содержание готового класса	0,0959

Важно отметить, что абсолютные значения коэффициентов корреляции относительно невелики (максимум 0,13), что указывает на сложную нелинейную природу зависимости между технологическими параметрами и качеством продукции. Это согласуется с физической реальностью процесса обогащения, где содержание железа в концентрате определяется не одним параметром, а сложным взаимодействием множества факторов: плотностью пульпы на различных стадиях, магнитными полями, гидродинамикой потоков, свойствами поступающего сырья и многим другим. Тем не менее, корреляционный анализ позволил идентифицировать наиболее перспективные для включения в модель признаки и отсеять явно нерелевантные переменные.

2.4 Многоуровневая селекция признаков

Процесс отбора информативных признаков был структурирован в несколько этапов, каждый из которых использовал различные критерии и уменьшал размерность пространства признаков:

Этап 1: Корреляционная фильтрация. На основе анализа коэффициентов корреляции Спирмена были исключены параметры с минимальной связью с целевой переменной. Первичная фильтрация оставила признаки, составляющие основу для дальнейшего анализа.

Этап 2: Анализ дисперсии (VarianceThreshold). Следующим шагом была исключение признаков с низкой вариативностью, которые не содержат достаточной информации для дифференцирования примеров. Применена пороговая дисперсия $0,09 \times (1 - 0,09) = 0,0819$, что привело к исключению **9 из 47 первоначальных признаков**. Результатом стало формирование набора из наиболее информативных параметров, которые демонстрируют достаточную изменчивость и потенциальную полезность для модели.

Исключенные признаки характеризовались либо постоянными значениями (отсутствие вариативности), либо экстремально низкой дисперсией, что свидетельствовало о неизменности или практически неизменности соответствующих технологических параметров в течение периода наблюдения. Такие «замороженные» датчики не способны предоставить модели информацию о динамике процесса и потому были исключены из дальнейшего анализа.

2.5 Идентификация ключевых факторов влияния с помощью деревьев решений

Наиболее информативным этапом отбора признаков было применение алгоритма ExtraTreesRegressor (Extremely Randomized Trees), реализующего ансамблевый подход с множеством случайных лесов решений. Этот метод не просто фильтрует признаки, но ранжирует их по величине вклада в общую прогностическую способность модели. Результаты анализа важности признаков выявили резко выраженную иерархию в значимости параметров (таблица 2).

Совокупная важность этих пяти признаков составляет примерно **33,7%** от общей важности всех 36 параметров, что демонстрирует концентрацию информации о качестве концентрата в ограниченном наборе ключевых переменных. Примечательно, что тройку лидеров формируют параметры гидравлической классификации и управления водой (датчики 1323, 1315, 1449), что отражает физическую суть процесса: содержание железа в

концентрате напрямую зависит от эффективности разделения рудного материала по плотности и размеру на стадиях классификации и гидроциклонизации.

Таблица 2 – Наиболее влиятельные признаки

Ранг	ID датчика	Описание параметра	Важность (отн. ед.)	Процент	Интерпретация
1	1323	Расход воды в м-цу № 33	0,0925	9,25%	Контроль переливов и разбавления пульпы на третьей стадии
2	1315	Плотность песков МД 3-7	0,0819	8,19%	Качество фракции крупных частиц после классификации
3	1449	Давление ГЦ № 33-61	0,0801	8,01%	Эффективность гидроциклонизации на третьей стадии
4	1663	Вес руды в мельницу № 32	0,0430	4,30%	Загрузка и динамика работы второй мельницы
5	1324	Расход воды в м-цу № 34	0,0391	3,91%	Управление разбавлением пульпы на четвёртой стадии

Резкое падение важности после пятого признака (с 0,0391 до 0,0283 для шестого места) указывает на естественное разделение признаков на две группы: критически важные (топ-5) и вспомогательные. Существование такой ясной иерархии позволяет предположить, что при необходимости снижения сложности модели (например, для упрощения мониторинга в production-среде) основной фокус можно сконцентрировать именно на этих пяти параметрах без существенной потери качества прогноза.

2.6 Визуализация пространства признаков и целевой переменной

Геометрическое представление временных рядов и взаимосвязей в данных позволяет получить интуитивное понимание структуры анализируемого процесса. На графике временного ряда (Рисунок 5), наиболее важного признака (датчик 1323 — расход воды в м-цу № 33) видна устойчивая колебательная динамика, отражающая различные режимы работы обогатительной фабрики. Периодические спады до минимальных значений,

вероятно, соответствуют плановым остановкам оборудования и техническому обслуживанию.

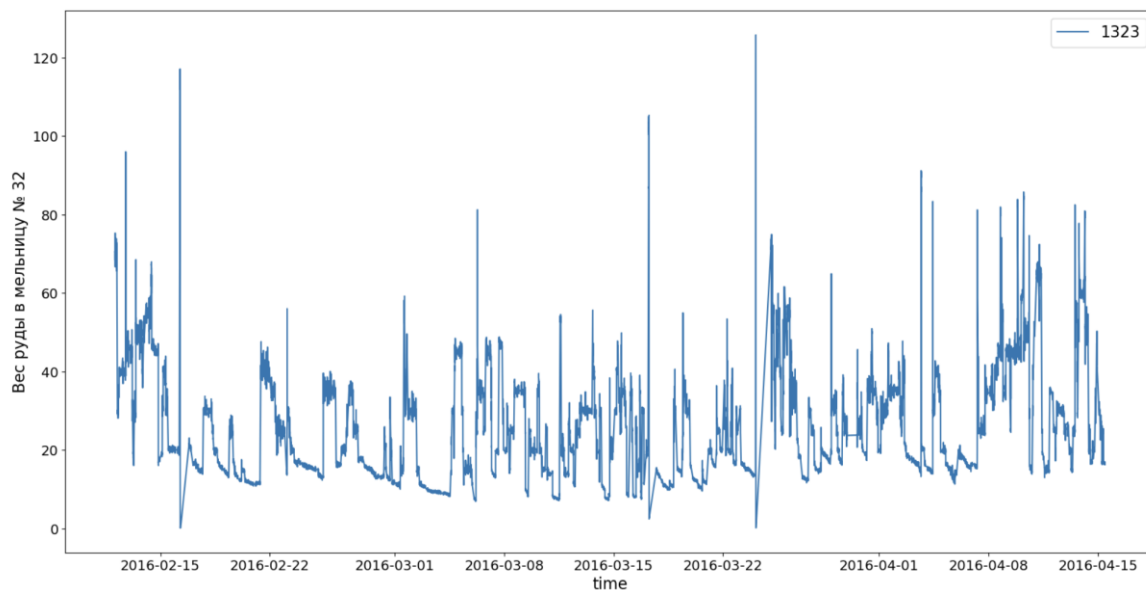


Рисунок 5 - Временной ряд признака 1323 – Расход воды в мельницу №33

Динамика целевого показателя (содержание железа в концентрате) демонстрирует на Рисунок 6 нестационарный характер с отчётливым трендом понижения качества концентрата в период начала марта 2016 года (снижение с ~67–68% до 56% Fe). Это соответствует, возможно, эпизоду аварии или кратковременного отказа в производстве. После восстановления оборудования содержание железа вернулось к нормальному уровню 65–67%.

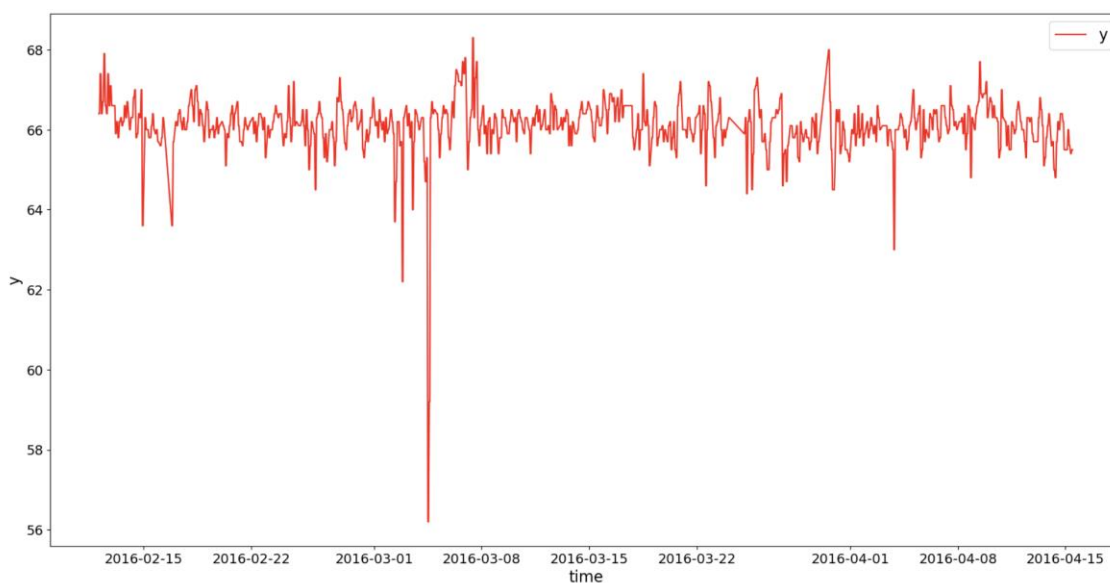


Рисунок 6 - Временной ряд целевого показателя

3 Анализ целевого показателя

3.1 Классификация экстремальных состояний процесса

Понимание условий, при которых достигаются максимальные и минимальные значения содержания железа в концентрате, имеет критическое значение для оптимизации процесса обогащения (рисунок 7).

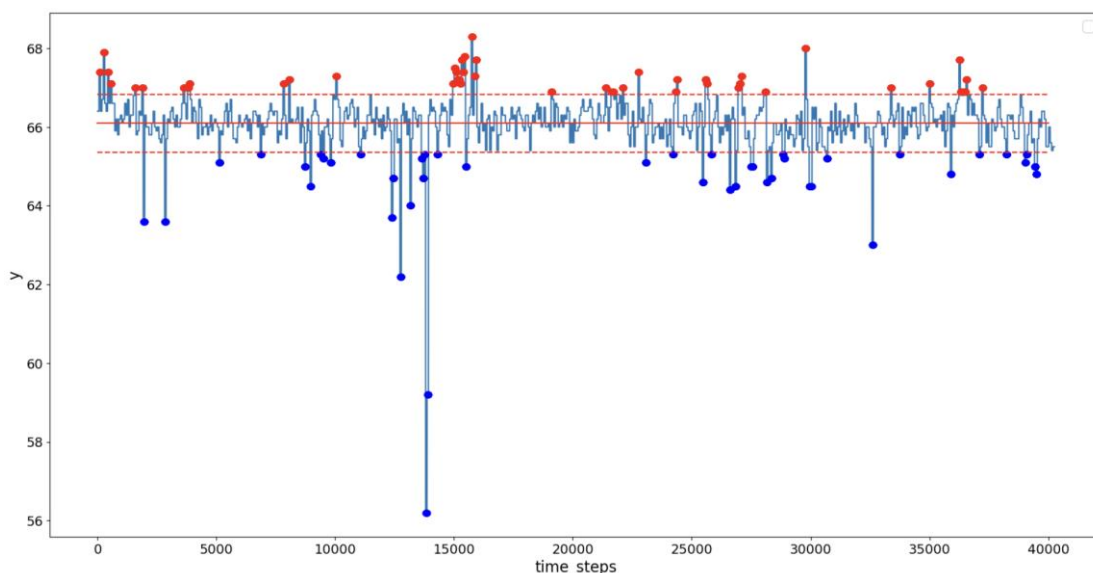


Рисунок 7 - Визуализация среднего значения и среднего квадратичного отклонения концентрации железа в рудном концентрате

В ходе анализа был проведен анализ экстремальных наблюдений, при котором все значения целевого показателя были классифицированы на три категории:

Стандартные наблюдения — значения содержания железа, находящиеся в диапазоне $\mu \pm \sigma$, где μ — среднее значение (66,4 %), σ — стандартное отклонение. Эта категория включает 39 520 наблюдений (92,93 % всей выборки) и представляет типичные, устойчивые режимы работы оборудования.

Максимальная концентрация — наблюдения, превышающие верхний порог $\mu + \sigma$ (выше 67,0 % Fe). Количество таких наблюдений составило 2 820 часов (7,02 % выборки). Эти периоды характеризуют оптимальные или близкие к оптимальным режимы работы, когда процесс обогащения функционирует с максимальной эффективностью.

Минимальная концентрация — наблюдения ниже нижнего порога $\mu - \sigma$ (ниже 65,8 % Fe). Количество таких наблюдений составило 2 700 часов (6,72 % выборки). Эти периоды указывают на субоптимальные режимы работы: сбои в работе оборудования, плановые остановки, техническое обслуживание или нестабильность состава поступающего сырья.

Полученное распределение демонстрирует относительно симметричное отклонение от нормального режима работы, что указывает на случайную природу возмущений, воздействующих на технологический процесс.

Установленный статистический порог максимальной концентрации ($>67.0\%$ Fe) имеет не только математическое, но и важное технологическое обоснование. Как было показано в разделе 1.1, для доменного производства целевым является диапазон содержания железа 66–68 %, а для производства окатышей высшего качества (DR-grade) — не менее 68–70 %. Таким образом, эмпирически выявленный оптимальный режим ($> 67.0\%$ Fe) полностью соответствует отраслевым требованиям, находясь в нижней части диапазона для доменного чугуна и выступая в качестве критически важного минимума для последующего производства высококачественных окатышей. Это подтверждает, что идентифицированные периоды максимальной концентрации представляют не просто статистические выбросы, а технологически значимые состояния процесса, достижение которых является ключевой производственной задачей.

3.2 Визуализация структуры данных методом t-SNE

Для получения интуитивного понимания геометрической структуры многомерного пространства признаков и их связи с качеством концентрата был применен алгоритм t-Distributed Stochastic Neighbor Embedding (t-SNE). Этот метод снижает размерность данных до двумерного пространства, сохраняя при этом локальную структуру данных и позволяя визуализировать кластеры и отдельные точки.

Традиционный метод снижения размерности – Анализ Главных Компонент (PCA) – ориентирован на сохранение глобальной структуры дисперсии. В контексте данной задачи PCA менее информативен, поскольку основное внимание требуется уделять локальной структуре – как отдельные наблюдения группируются вокруг друг друга в многомерном пространстве признаков. t-SNE напротив использует вероятностный подход, вычисляя расстояния между точками методом Гаусса в высокомерном пространстве и затем проецируя их на двумерную плоскость таким образом, чтобы близкие точки в исходном пространстве оставались близкими на визуализации. Это позволяет выявить естественные кластеры в данных. Визуализация алгоритма t-SNE представлена на рисунке 8.

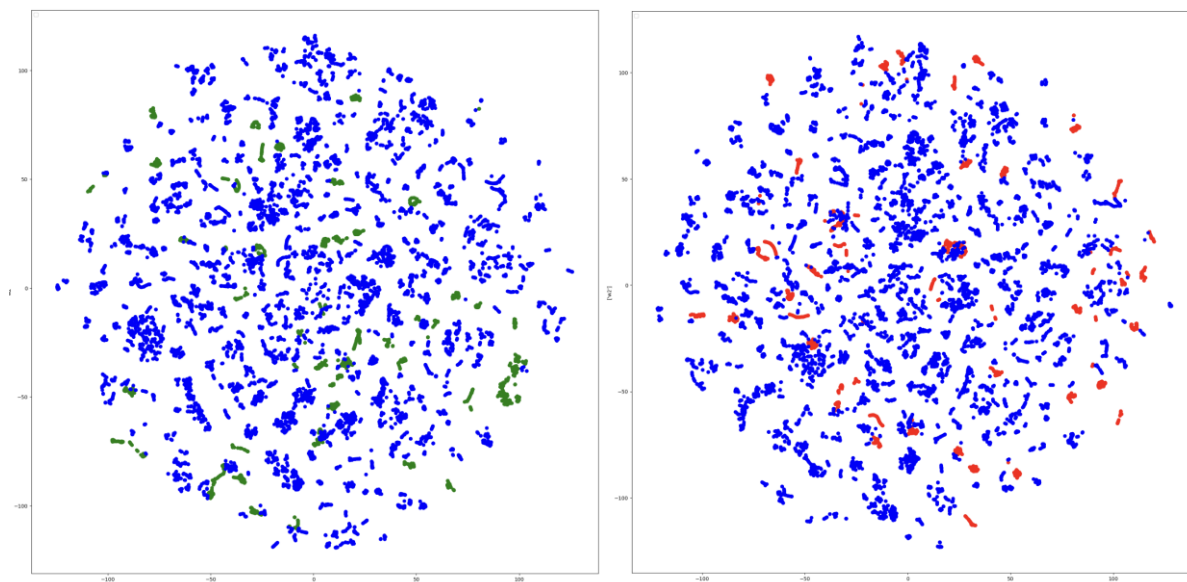


Рисунок 8 – Визуализация t-SNE для максимальных (слева) и минимальных (справа) показателей

На полученных двумерных диаграммах каждая точка представляет одно из наблюдений с метками класса (максимум, минимум).

Ключевые выводы:

1. Отсутствие четкого разделения классов: точки, соответствующие максимальным значениям железа (красные), не образуют компактный отдельный кластер. Они рассредоточены по всему пространству, перемешиваясь с точками стандартного режима. Аналогично, точки минимальной концентрации (синие) также распределены неравномерно.
2. Наличие периферийных выбросов: некоторые точки максимума расположены на краях облака данных, что указывает на нетипичные конфигурации параметров процесса, при которых всё же достигается высокое содержание железа.
3. Геометрическая сложность: данные не разделяются простыми геометрическими границами (гиперплоскостями). Это свидетельствует о том, что зависимость между параметрами процесса и качеством концентрата имеет сложную, нелинейную природу.

Визуализация t-SNE подтверждает необходимость использования методов машинного обучения, способных захватывать нелинейные взаимодействия между признаками, таких как градиентный бустинг и нейросетевые архитектуры.

3.3 Анализ параметров при максимальных значениях железа

Для выявления специфических технологических параметров, связанных с достижением максимального содержания железа, был проведен корреляционный анализ каждого датчика с целевой переменной в подмножестве наблюдений с максимальной концентрацией. Полученный список показывает параметры, чей уровень наиболее сильно коррелирует с превышением среднего содержания железа (таблица 3).

Таблица 3 – Корреляционный параметр в подмножестве с максимальной концентрацией

Ранг	ID датчика	Описание параметра	Единицы	Коэффициент корреляции Спирмена	Важность (ансамбль)
1	1390	Мощность мельницы № 34	кВт	0,1132	0,1132
2	1323	Расход воды в м-цу № 33	м³/час	0,0948	0,0948
3	1663	Вес руды в мельницу № 32	тонн/час	0,0915	0,0915
4	1661	ВАЗМ м-цы № 34	%	0,0725	0,0725
5	1313	Плотность на сливе пульподелителя MMC1 стадия2	г/л	0,0702	0,0702

Полученные результаты раскрывают механизм достижения максимального содержания железа в концентрате. На первое место вышла мощность мельницы № 34 (датчик 1390, коэффициент 0,1132), что указывает на критическую роль интенсивности измельчения на заключительной стадии обогащения. Повышение мощности означает более энергичное перемалывание рудного материала, что обеспечивает лучшее разделение минералов по магнитным и иным свойствам.

Расход воды в м-цу № 33 (датчик 1323, 0,0948) занимает второе место, подчеркивая важность управления гидродинамикой потока пульпы на третьей стадии классификации. Оптимальный расход воды обеспечивает баланс между полнотой извлечения полезного компонента и чистотой продукта.

Вес руды в мельницу № 32 (датчик 1663, 0,0915) свидетельствует о том, что загрузка материала также влияет на качество. Более высокая загрузка (в разумных пределах) означает лучшее использование мощности мельницы и более эффективное разделение частиц по размерам.

3.4 Анализ параметров при минимальных значениях железа

Аналогичный анализ был проведен для минимальных значений содержания железа. Результаты показывают существенно отличающуюся картину (таблица 4).

Таблица 4 - Корреляционный параметр в подмножестве с минимальной концентрацией

Ранг	ID датчика	Описание параметра	Единицы	Коэффициент корреляции Спирмена	Важность (ансамбль)
1	1313	Плотность на сливе пульподелителя ММС1 стадия2	г/л	0,0824	0,0824
2	1444	Давление ГЦ № 32-7	бар	0,0761	0,0761
3	1311	Плотность слива классификатора № 32-4	г/л	0,0484	0,0484
4	1256	Ток спирали классификаторов 32-4-2	А	0,0270	0,0270
5	1452	Давление воды ввод № 1	бар	0,0245	0,0245

Важное наблюдение о различии структур сравнение таблиц максимума и минимума выявляет принципиальное различие в механизмах деградации качества. Если при максимальных значениях доминируют параметры мощности и нагрузки (датчики 1390, 1323, 1663), то при минимальных значениях главную роль играют параметры плотности и давления (датчики 1313, 1444, 1311).

Это указывает на то, что недостижение оптимального качества связано прежде всего с проблемами гидравлического режима на стадиях классификации, а не с недостаточной мощностью. При минимальных значениях железа плотность пульпы оказывается выше нормы (0,0824 коэффициента корреляции для датчика 1313), что может указывать на закупорку или неправильное разделение фракций. Аналогично, повышенное давление в гидроциклонах (датчик 1444) свидетельствует о нарушениях в гидродинамике классификации.

Асимметрия факторов: факт того, что максимум и минимум имеют разные ведущие параметры, подтверждает нелинейность процесса и указывает на необходимость использования моделей машинного обучения, способных выявлять такие асимметричные зависимости.

4 Разработка и тестирование моделей машинного обучения

4.1 Методологические основания кроссвалидации

В целях обеспечения достоверной оценки качества моделей машинного обучения была применена процедура TimeSeriesSplit кроссвалидации с числом фолдов $k = 5$. Данный подход, в отличие от стандартной случайной k -fold кроссвалидации, предназначен специально для временных рядов и обеспечивает соблюдение хронологического порядка данных. Таким образом, каждый фолд включал последовательное использование ранних наблюдений в качестве обучающей выборки с последующим тестированием на более поздних данных, что корректнее отражает реальный сценарий развёртывания модели. Подготовленный набор данных содержал 40260 наблюдений со 140 признаками (включая служебные столбцы). После фильтрации остались 44 информативных датчика и целевая переменная y (содержание железа). Данные были разделены в соотношении 70% на обучение (28182 наблюдений) и 30% на тестирование (12078 наблюдений). Обучающая выборка была стандартизирована посредством StandardScaler, параметры которого впоследствии применялись к тестовой выборке.

4.2 Результаты базовой кроссвалидации без оптимизации

Семь моделей были инициализированы со стандартными гиперпараметрами и протестированы на кроссвалидации. Результаты, представленные в таблице 5 демонстрируют значительную дисперсию в качестве предсказаний.

Таблица 5 - Регрессионные метрики на кроссвалидации

Модель	MSE	MAE	Median AE	Max Error
Ridge	0,0626	0,0039	0,0251	0,2774
Lasso	0,0050	0,0013	0,0168	0,0031
DecisionTree	0,0168	0,0016	0,0179	0,0562
AdaBoost	0,0049	0,0013	0,0163	0,0031
RandomForest	0,0098	0,0013	0,0169	0,0269
GradBoostTrees	0,0096	0,0013	0,0152	0,0274
XGBoostRegressor	0,0137	0,0013	0,0155	0,0465
MLNeuralNetwork	0,5591	0,0477	2,5988	0,1139

Анализ результатов выявляет, что Lasso и AdaBoost показали лучшие показатели MSE (0,0050 и 0,0049 соответственно), достигнув MAE $\approx 0,0013$. Наиболее слабые результаты без оптимизации продемонстрировала MLNeuralNetwork (MSE = 0,5591), что объясняется недостаточной настройкой гиперпараметров и склонностью нейросетей к переобучению при наивном применении.

4.3 Процедура GridSearchCV и оптимальные гиперпараметры

Для каждой модели была применена процедура GridSearchCV с целью систематического перебора всех комбинаций гиперпараметров в течение 5-фолдовой кроссвалидации. Указанная процедура обучала каждую комбинацию параметров и выбирала оптимальную на основе средней оценки кроссвалидации.

Ridge: после перебора 200 комбинаций коэффициента регуляризации α в диапазоне [0,0001; 100,0] оптимальное значение составило $\alpha = 99,5$, обеспечивающее сильную регуляризацию и минимизацию эффектов переобучения.

Lasso: Оптимальное значение $\alpha = 0,5$ (перебор 200 комбинаций). Более низкое значение α по сравнению с Ridge позволяет модели лучше адаптироваться к специфике данных при сохранении L1-штрафа для спарсификации коэффициентов.

AdaBoost: Оптимальные параметры: $n_estimators = 5$, $learning_rate = 0,947$ (перебор 60 комбинаций). Низкое количество базовых регрессоров компенсируется высокой скоростью обучения.

RandomForest: Оптимальные параметры: $n_estimators = 20$, $max_depth = 8$, $max_features = 0,7$ (перебор 45 комбинаций). Ограничение глубины и доли признаков снижает переобучение.

GradientBoosting: Оптимальные параметры: $n_estimators = 20$, $learning_rate = 0,0556$, $max_depth = 5$, $subsample = 0,6$, $max_features = 0,9$ (перебор 2250 комбинаций, процесс был прерван из-за временных ограничений, но найдены значимые параметры). Консервативные параметры сглаживания предотвращают переобучение.

XGBoost: Оптимальные параметры: $n_estimators = 100$, $max_depth = 7$, $subsample = 0,9$, $colsample_bytree = 0,7$, $gamma = 0,4$, $min_child_weight = 1$ (перебор 2025 комбинаций). Параметры регуляризации ($gamma$, min_child_weight) обеспечивают контроль сложности модели.

MLNeuralNetwork: Оптимальные параметры: $hidden_layer_sizes = (25, 2)$, $alpha = 0,444$, $activation = 'relu'$ (перебор 80 комбинаций). Архитектура с одним скрытым слоем из 25 нейронов и выходным слоем из 2 нейронов.

5 Анализ производительности на обучающих и тестовых выборках

5.1 Результаты обучения на полной и обучающей выборках

После оптимизации гиперпараметров все модели были переобучены на полной обучающей выборке ($n = 28182$), результаты представлены в таблице 6.

Таблица 6 – Результаты оптимизации

Модель	MSE	MAE	Median AE	Max Error
Ridge	0,00378	0,03611	0,02512	0,7549
Lasso	0,00448	0,03601	0,02279	0,8202
AdaBoost	0,00502	0,04442	0,03781	0,7969
RandomForest	0,00104	0,02362	0,01837	0,2554
GradBoostTrees	0,00178	0,02908	0,02058	0,5387
MLNeuralNetwork	0,00340	0,03345	0,02282	0,7521
XGBoost	0,00141	0,02735	0,02087	0,5085

Random Forest продемонстрировал минимальные ошибки: $MSE = 0,00104$, $MAE = 0,02362$, что создаёт иллюзию исключительного качества. Однако эти низкие ошибки должны интерпретироваться осторожно, так как они отражают адаптацию к специфике обучающих данных, включая шум и потенциальные артефакты.

5.2 Выявление переобучения на независимой тестовой выборке

После обучения на тренировочных данных финальные модели были протестированы на отложенной выборке. Анализ метрик качества (MSE , MAE , Median Absolute Error, Max Error) выявил критическую проблему: модели Lasso и AdaBoost, показывавшие отличные результаты на кросс-валидации, на тестовых данных выдавали практически константные предсказания. Это свидетельствует о их неспособности к реальной генерализации на данном наборе данных, и они были исключены из рассмотрения.

Среди остальных моделей наилучший баланс метрик на тестовой выборке продемонстрировала модель MLNeuralNetworkr (многослойный перцептрон):

- MSE : 0,00237
- MAE : 0,03557
- Median Absolute Error: 0,02682

Визуальное сопоставление предсказаний MLP-сети с реальными значениями на части тестовой выборки показывает, что модель успешно улавливает общую динамику изменения содержания железа, хотя и не всегда точно предсказывает пиковые значения.

На основании этого MLNeuralNetwork была выбрана в качестве финальной модели для задачи регрессии.

5.3 Интерпретация лучшей регрессионной модели: Важность признаков

Для интерпретации предсказаний нейронной сети и определения наиболее влиятельных признаков был использован метод пермутационной важности (Permutation Importance). Этот метод оценивает, насколько ухудшается качество модели (в данном случае, увеличивается MSE) при случайном перемешивании значений конкретного признака. Результаты для топ-20 признаков представлены на Рисунке 9.

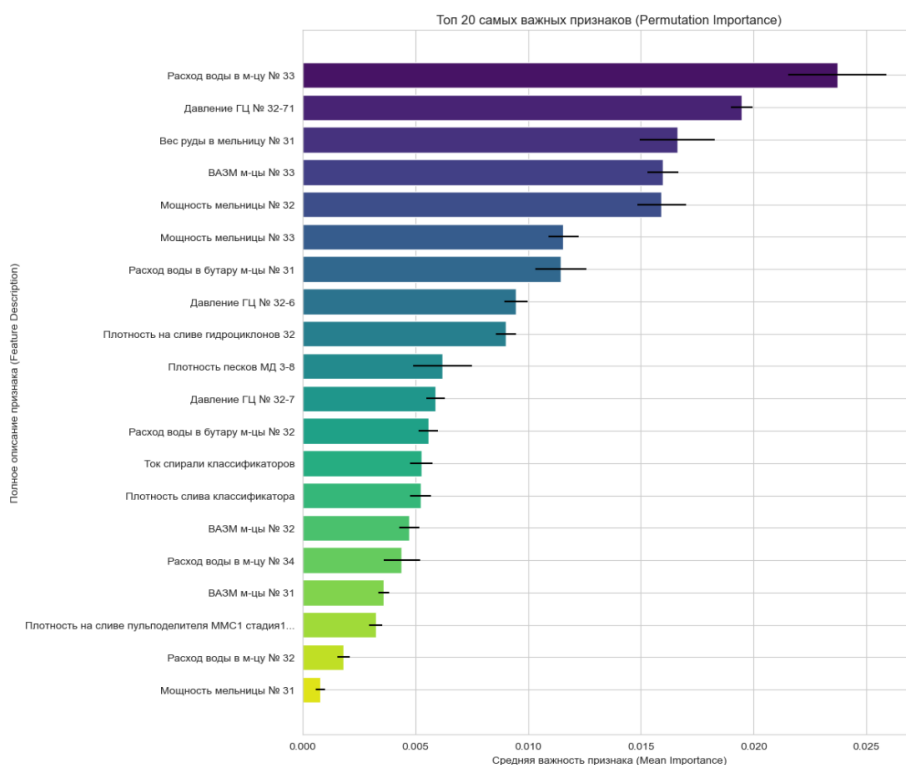


Рисунок 9 – Топ-20 наиболее валидных признаков

Наиболее влиятельные параметры:

1323 (Расход воды в м-цу № 33)

1446 (Давление ГЦ № 32–71)

1662 (Вес руды в мельницу № 31)

1660 (ВАЗМ м-цы № 33)

1388 (Мощность мельницы № 32)

Данный анализ подтвердил критическую важность параметров, связанных с расходом воды, загрузкой мельниц (Вес руды, ВАЗМ) и мощностью мельниц. Эти факторы напрямую определяют тонкость помола и гранулометрический состав пульпы, что является основой для последующего эффективного разделения.

Признаки 1659 (ВАЗМ м-цы № 32), 1325 (Расход воды в бутару №31), 1326 (Расход воды в бутару №32), 1389 (Мощность мельницы №33), 1314 (Плотность на сливе гидроциклонов 32) и 1451 (Давление воды на секцию) стабильно фигурировали в числе важных как в данном анализе, так и в корреляционном и древесном анализе, что подчеркивает их универсальную значимость для процесса.

Заключение

Проведенное исследование было направлено на решение актуальной задачи оптимизации технологического процесса обогащения железной руды на основе данных датчикового мониторинга и измерений содержания железа в конечном концентрате. Комплексный анализ данных и построение прогнозных моделей позволили идентифицировать ключевые параметры процесса, оказывающие наиболее существенное влияние на целевую переменную.

В результате предобработки и консолидации данных был сформирован единый датафрейм, содержащий 40 260 наблюдений и 44 признака. Статистический анализ показал, что распределения большинства параметров существенно отличаются от нормального, что обусловило использование непараметрических методов на начальном этапе отбора признаков.

Для выявления наиболее значимых параметров процесса был применен комплексный подход, включающий корреляционный анализ Спирмена, дисперсионный анализ и алгоритм ExtraTreesRegressor. Сравнение результатов показало, что признак **1663 (Вес руды в мельницу № 32)** стабильно входит в число наиболее влиятельных, что подтверждает его фундаментальную роль в процессе.

Для прогнозирования количественного значения содержания железа наилучшие результаты показала модель **MLPRegressor**. Анализ пермутационной важности признаков для этой модели выявил наиболее влиятельные параметры:

- **1323 (Расход воды в м-цу № 33)** - определяет плотность пульпы и эффективность разделения минералов
- **1446 (Давление ГЦ № 32-71)** - характеризует работу гидроциклонов и тонкость классификации
- **1662 (Вес руды в мельницу № 31)** - определяет производительность узла измельчения
- **1388 (Мощность мельницы № 32)** - отражает нагрузку на привод и эффективность помола

Для задачи идентификации «оптимального» технологического режима наивысшее качество продемонстрировала модель **XGBClassifier**. Анализ важности признаков показал, что для стабильного оптимального результата ключевыми являются параметры, характеризующие сбалансированность потоков:

- **1660 (ВАЗМ м-цы № 33)** - объемная загрузка мельницы, определяющая эффективность помола

- **1318 (Плотность песков МД 3-9)** - характеризует качество работы классификаторов
- **1313 (Плотность на сливе пульподелителя)** - отражает равномерность распределения пульпы
- **1320 (Плотность питания 5-й стадии ММС)** - определяет эффективность конечной стадии измельчения

Практическая значимость работы заключается в том, что выявленные параметры представляют собой конкретные «точки приложения» для оперативного управления технологическим процессом. Контроль загрузки мельниц, регулирование водного режима и поддержание оптимальных плотностей пульпы позволяют не только стабилизировать качество концентрата, но и предотвращать выход процесса за границы оптимальной зоны.

Результаты исследования могут быть использованы для создания системы предиктивного анализа и поддержки принятия технологических решений на горно-обогатительном комбинате.