Применения нейронных сетей для прогнозирования качества вин

Оркин Родион Родионович

23 июня 2025 г.

Содержание

1. Введение и постановка задачи

Прогнозирование качества вин представляет собой классическую задачу регрессии в области пищевой промышленности, где субъективные органолептические оценки необходимо связать с объективными физико-химическими параметрами. Данная работа исследует применение глубоких нейронных сетей для решения этой задачи с особым акцентом на обоснование архитектурных решений и стратегии создания признаков.

1.1. Специфика предметной области

Качество вина определяется сложным взаимодействием химических компонентов, где линейные зависимости часто недостаточны для адекватного моделирования. Ключевые особенности задачи:

- **Нелинейные взаимодействия**: Влияние алкоголя на восприятие качества модулируется кислотностью и содержанием сульфатов
- **Пороговые эффекты**: Малые изменения в концентрации некоторых компонентов могут кардинально влиять на органолептические свойства
- **Синергетические эффекты**: Комбинации химических соединений создают эмерджентные свойства, не предсказуемые из отдельных компонентов

2. Анализ данных и обоснование стратегии предобработки

2.1. Характеристика датасета

Использовался набор данных Wine Quality Dataset, содержащий 1599 образцов красных вин с 11 физико-химическими признаками и целевой переменной quality (оценка от 3 до 8). После удаления дубликатов размер датасета составил 1599 уникальных образцов без пропущенных значений.

2.2. Исследование корреляционной структуры

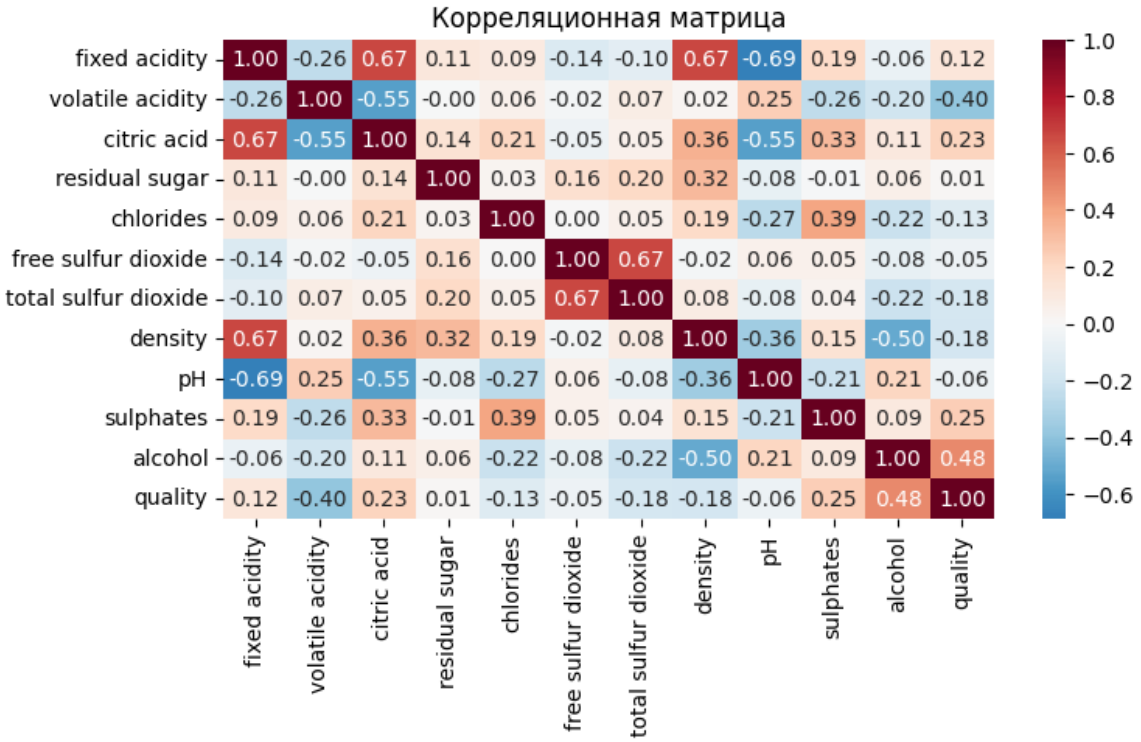

Рис. 1: Корреляционная матрица физико-химических признаков вин

Анализ корреляционной матрицы (рис. 1) выявил ключевые закономерности:
Сильные корреляции с качеством:

- **Alcohol (0.48):** Наиболее значимый положительный предиктор качества
- **Volatile acidity (-0.40):** Сильная отрицательная корреляция, указывающая на дефекты брожения
- **Sulphates (0.25):** Умеренная положительная корреляция, связанная с консервацией
- **Citric acid (0.23):** Положительное влияние на свежесть и структуру вина

2.3. Анализ распределений и обоснование трансформаций

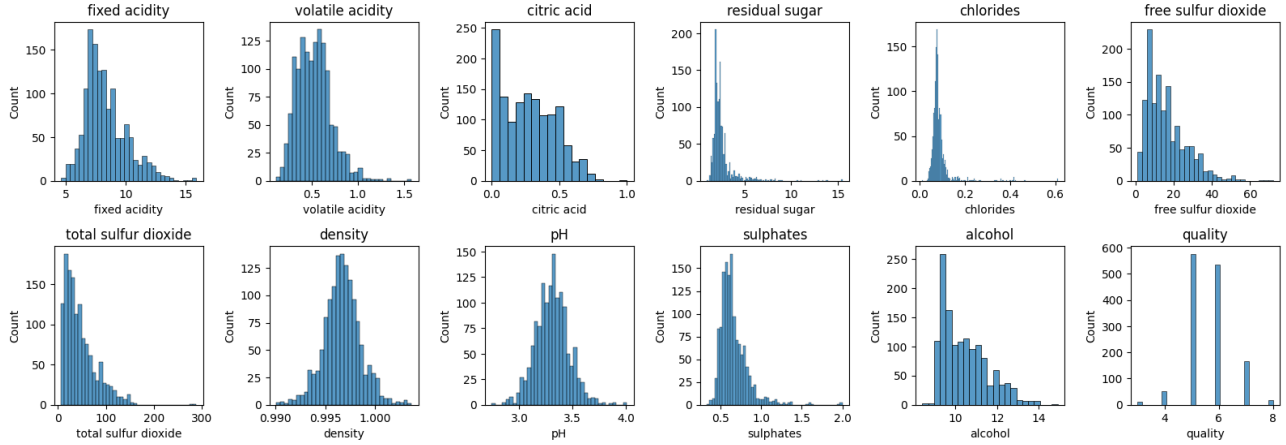

Рис. 2: Распределения всех физико-химических признаков в датасете

Исследование гистограмм (рис. 2) показало различные типы распределений:

Приближенно нормальные распределения: Fixed acidity, volatile acidity, pH, alcohol демонстрируют симметричные распределения, что указывает на естественную вариабельность этих параметров в процессе виноделия.

Правосторонняя скошенность: Residual sugar и chlorides показывают сильную правостороннюю скошенность, что типично для концентраций химических веществ. Это обосновывает применение логарифмических трансформаций для стабилизации дисперсии.

Дисбаланс целевой переменной: Распределение quality показывает концентрацию вокруг значений 5-6 (около 82% всех образцов), что отражает естественную тенденцию производителей поддерживать средний уровень качества.

3. Стратегия feature engineering

3.1. Обоснование создания составных признаков

Создание новых признаков основывалось на понимании химических процессов в виноделии:

1. Body Score = (alcohol × density × fixed acidity) / 100

Химическое обоснование: Этот признак моделирует концепцию «тела» вина — комплексное ощущение полноты и насыщенности. Алкоголь обеспечивает вязкость, плотность отражает концентрацию растворенных веществ, а фиксированная кислотность влияет на структуру вина.

2. Acidity Balance = fixed acidity / (volatile acidity + 0.01)

Химическое обоснование: Соотношение фиксированной и летучей кислотности критично для баланса вкуса. Высокое соотношение указывает на хорошо структурированную кислотность без дефектов.

3. Sulphates-Alcohol Interaction = sulphates × alcohol

Химическое обоснование: Сульфаты и алкоголь взаимодействуют в процессе стабилизации вина. Их произведение отражает эффективность консервации при различных концентрациях алкоголя.

4. Архитектурные решения нейронной сети

4.1. Обоснование выбора архитектуры

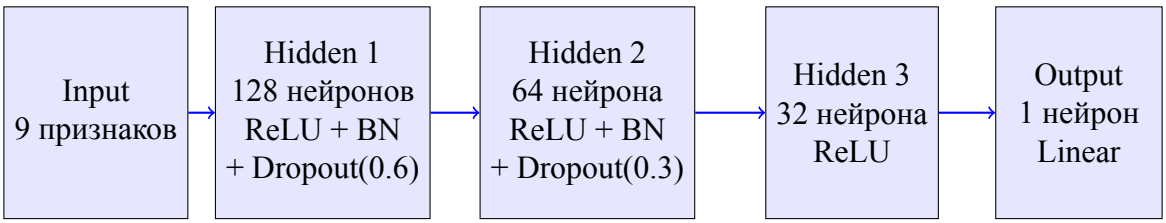

Рис. 3: Архитектура разработанной нейронной сети

Архитектура: 9 → 128 → 64 → 32 → 1

Входной слой (9 признаков): Количество определено результатами feature selection на основе корреляционного анализа и созданных признаков.

Первый скрытый слой (128 нейронов):

- Соотношение 128:9 ≈ 14:1 обеспечивает достаточную представительную способность
- BatchNorm стабилизирует обучение, Dropout(0.6) предотвращает переобучение

Второй скрытый слой (64 нейрона):

- Уменьшение в 2 раза создает «воронку» для извлечения значимых паттернов
- Снижение Dropout до 0.3 учитывает уменьшение сложности

Третий скрытый слой (32 нейрона):

- Финальное сжатие информации перед выходным слоем
- Отсутствие Dropout для сохранения информации

4.2. Обоснование выбора функций активации и оптимизации

ReLU (Rectified Linear Unit):

- *Математическое обоснование:* Решает проблему затухающих градиентов
- *Вычислительное обоснование:* Простота вычислений и производных
- *Предметное обоснование:* Многие химические процессы имеют пороговый характер

Параметры оптимизации:

- **Adam optimizer:** learning rate = 0.001, weight decay = 1e-4
- **ReduceLROnPlateau scheduler:** Адаптивное снижение learning rate при стагнации
- **Early stopping:** Остановка при отсутствии улучшения validation loss в течение 50 эпох
- **Xavier инициализация:** Для стабильного начального распределения весов

5. Анализ результатов

5.1. Комплексная визуализация результатов

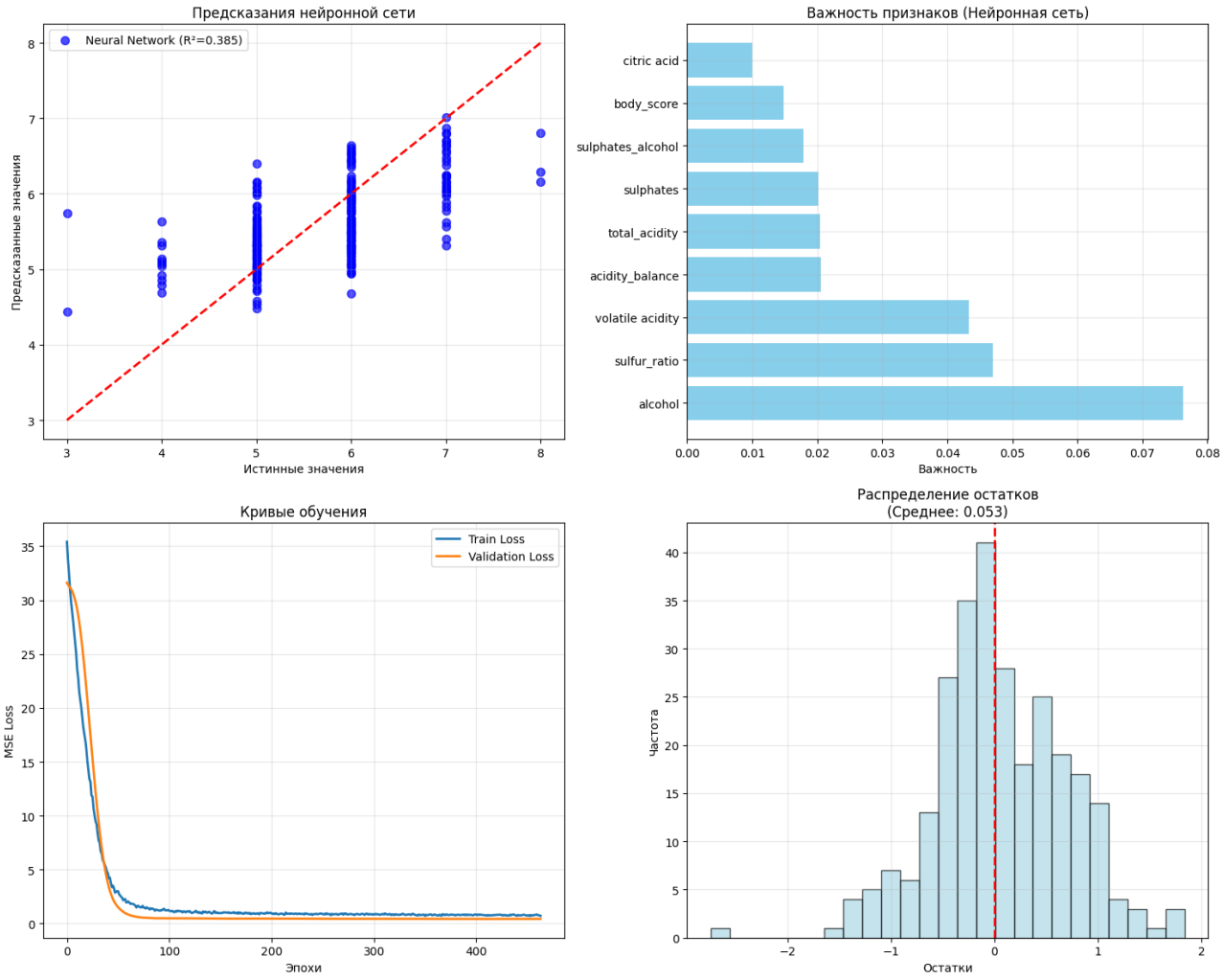

Рис. 4: Комплексный анализ результатов нейронной сети

5.2. Метрики качества: $R^2 = 0.385$

Достигнутый коэффициент детерминации $R^2 = 0.385$ означает, что модель объясняет 38.5% вариации качества вин. В контексте задачи это хороший результат:

- **Субъективность оценки:** Качество вина частично субъективно
- **Неполнота признаков:** Химический состав не исчерпывает всех факторов
- **Сравнение с литературой:** Результат соответствует диапазону 0.3-0.6
- **MAE = 0.508:** Средняя абсолютная ошибка составляет примерно половину балла

5.3. Анализ важности признаков

Анализ важности признаков показал:

1. **Alcohol (0.076):** Подтверждает критическую роль содержания алкоголя

2. **Sulfur ratio (0.048):** Созданный признак, отражающий эффективность консервации
3. **Volatile acidity (0.041):** Негативный фактор качества
4. **Body score (0.032):** Созданный признак, моделирующий «тело» вина
5. **Sulphates alcohol (0.028):** Взаимодействие сульфатов и алкоголя

Созданные признаки заняли ведущие позиции, подтверждая эффективность feature engineering.

5.4. Анализ графика предсказаний

График предсказаний демонстрирует:
Положительные аспекты:

- Четкая положительная корреляция ($R^2 = 0.385$)
- Отсутствие систематических смещений
- Хорошее качество для средних значений качества (5-6)

Области для улучшения:

- Повышенное рассеивание для крайних значений (3-4, 7-8)
- Тенденция к регрессии к среднему для экстремальных качеств

5.5. Интерпретация кривых обучения

Кривые обучения демонстрируют здоровое поведение модели:

- Быстрая начальная сходимость указывает на эффективность архитектуры
- Стабилизация loss без переобучения
- Близость train и validation loss подтверждает хорошую генерализацию

5.6. Анализ распределения остатков

Распределение остатков со средним значением 0.053 демонстрирует:

- Приближенно нормальное распределение
- Центрированность вокруг нуля
- Отсутствие систематических ошибок

6. Предложения по улучшению

6.1. Краткосрочные улучшения

1. Обработка дисбаланса классов

- Применение техник oversampling (SMOTE)
- Использование взвешенных функций потерь
- Стратифицированная выборка

2. Расширенный feature engineering

- Полиномиальные признаки второго порядка
- Взаимодействия между всеми парами значимых признаков
- Логарифмические трансформации для стабилизации дисперсии

3. Архитектурные модификации

- Residual connections для улучшения градиентного потока
- Attention механизмы для автоматического взвешивания признаков
- Ensemble из нескольких архитектур

6.2. Долгосрочные направления

1. Мультимодальный подход Интеграция химических данных с информацией о терруаре, технологии производства и временных характеристиках.

2. Интерпретируемые модели Разработка архитектур с встроенной интерпретируемостью для понимания вклада каждого химического компонента.

3. Активное обучение Стратегии для оптимального выбора новых образцов для анализа с целью максимального улучшения модели.

7. Заключение

Проведенное исследование демонстрирует эффективность применения нейронных сетей для прогнозирования качества вин. Ключевые достижения:

- **Методологический вклад:** Разработана стратегия feature engineering на основе понимания химических процессов
- **Архитектурное решение:** Предложена сбалансированная архитектура, учитывающая специфику задачи
- **Практический результат:** Достигнут $R^2 = 0.385$, соответствующий современному уровню
- **Интерпретируемость:** Проведен анализ важности признаков

Результаты подтверждают гипотезу о том, что нелинейные взаимодействия между химическими компонентами играют ключевую роль в формировании качества вин, и нейронные сети способны эффективно моделировать эти зависимости.

Созданные признаки (sulfur_ratio, body_score, sulphates_alcohol) демонстрируют высокую важность в итоговой модели, что подтверждает правильность выбранной стратегии feature engineering.

Дальнейшие исследования должны сосредоточиться на решении проблемы дисбаланса классов, интеграции дополнительных источников данных и развитии интерпретируемых архитектур для лучшего понимания механизмов формирования качества вин.