Итоговый отчёт

Тема: Проведение исследований с различными алгоритмами машинного обучения (KNN, Linear/Logistic, Decision Tree, Random Forest, Gradient Boosting) на двух задачах: регрессия и классификация

Автор: Белякова Софья Андреевна (М8О-407Б-22) **Дата: 26.12.2025**

ЛАБОРАТОРНАЯ РАБОТА №1 — (была описана в README в репозитории)

ЛАБОРАТОРНАЯ РАБОТА №2

Линейная и логистическая регрессия

1. Цель работы

Целью работы являлось исследование линейной и логистической регрессии, их возможностей и ограничений в задачах регрессии и классификации, а также сравнение базовых и улучшенных реализаций.

2. Линейная регрессия (регрессия)

Базовая линейная регрессия была применена к задаче прогнозирования арендной платы.

Результаты:

- R^2 на тестовой выборке около 0.50;
- высокая дисперсия ошибок;
- наблюдается нестабильность (test R^2 > train R^2).

Интерпретация:
Модель ограничена предположением линейной зависимости между признаками и целевой переменной.

После обработки выбросов качество улучшилось ($R^2 \approx 0.56$), а ошибки снизились.

3. Логистическая регрессия (классификация)

Модель применялась для прогнозирования успеваемости студентов.

Особенности:

- высокая интерпретируемость коэффициентов;
- чувствительность к масштабированию;
- зависимость качества от параметра регуляризации.

Оптимизация порога классификации позволила повысить F1-score и accuracy при дисбалансе классов.

4. Собственные реализации

Были реализованы собственные версии линейной и логистической регрессии.

Выводы:

- собственные модели корректно воспроизводят sklearn;
- различия в метриках находятся в пределах численной погрешности;
- улучшенный препроцессинг оказывает больший эффект, чем выбор реализации.

5. Выводы по лабораторной работе №2

Регрессионные модели просты и интерпретируемы, но чувствительны к качеству данных. Предобработка и работа с выбросами являются ключевыми факторами улучшения качества.

ЛАБОРАТОРНАЯ РАБОТА №3

Решающее дерево

1. Цель работы

Исследование алгоритма решающего дерева, его интерпретируемости, устойчивости и чувствительности к данным.

2. Базовый бейзлайн

Классификация:

- accuracy около 85%;
- модель устойчива при малой глубине.

Регрессия:

- R^2 около 0.34;
- высокая чувствительность к выбросам.

3. Улучшенный бейзлайн

После агрессивной предобработки:

- accuracy выросла до 0.92;
- R^2 регрессии достиг 0.67;
- RMSE и MAE снизились в разы.

EDA позволил выделить ключевые признаки, согласующиеся с предметной областью.

4. Собственная реализация

Выводы:

- GridSearch эффективно контролирует переобучение;
- качество сопоставимо и местами выше sklearn;
- интерпретируемость остаётся высокой.

5. Выводы по лабораторной работе №3

Решающее дерево является сильным интерпретируемым алгоритмом, однако требует обязательной предобработки данных.

ЛАБОРАТОРНАЯ РАБОТА №4

Случайный лес

1. Цель работы

Исследование ансамблевого метода Random Forest и его устойчивости к шуму и переобучению.

2. Бейзлайн

Без предобработки:

- классификация показывает завышенную accuracy;
- регрессия демонстрирует отрицательный R^2;
- наблюдается сильное переобучение.

3. Улучшенный бейзлайн

После очистки данных:

- R^2 регрессии ≈ 0.70;
- модель стала самой стабильной среди всех;
- важности признаков согласуются с предметной областью.

4. Ограничения

- модель плохо работает с миноритарным классом;
- оптимизация порога без балансировки неэффективна.

5. Выводы по лабораторной работе №4

Random Forest является одной из самых устойчивых моделей, однако качество данных критически влияет на результат.

ЛАБОРАТОРНАЯ РАБОТА №5

Градиентный бустинг и итоговое сравнение

1. Цель работы

Исследование градиентного бустинга и подведение итогов по всем алгоритмам.

2. Результаты

Классификация:

- высокая accuracy и F1-score;
- наблюдается переобучение;
- чувствительность к дисбалансу.

Регрессия:

- $R^2 \approx 0.36$;
- сильная зависимость от выбросов.

3. Итоговое сравнение алгоритмов

Классификация:

- все модели показывают высокую accuracy;
- качество определяется предобработкой, а не алгоритмом.

Регрессия:

- лучшим оказался Random Forest;
- RMSE более информативен, чем R^2.

4. Общие выводы

- качество данных важнее сложности модели;
- улучшенный бейзлайн — обязательный этап;
- ансамблевые методы наиболее устойчивы;
- собственные реализации имеют высокую образовательную ценность.

Выводы

В ходе выполнения лабораторных работ №1–5 был проведён комплексный сравнительный анализ алгоритмов машинного обучения для задач классификации и регрессии, включая KNN, логистическую и линейную регрессию, решающее дерево, случайный лес и градиентный бустинг. В задаче классификации все рассмотренные модели продемонстрировали высокие значения Accuracy (порядка 0.92–0.93), однако наличие выраженного дисбаланса классов привело к тому, что такие метрики, как Accuracy и F1-score, в ряде случаев переоценивали реальное качество моделей. Было установлено, что ни один из алгоритмов не решает проблему дисбаланса автоматически, а итоговое качество в большей степени определяется качеством предобработки данных и корректным выбором метрик, чем конкретным методом обучения.

В задаче регрессии выявлено, что целевая переменная характеризуется высокой вариативностью и наличием выбросов, вследствие чего многие модели без предварительной очистки данных демонстрировали отрицательные значения R^2, а метрика RMSE оказалась более информативной для оценки качества. Сравнение алгоритмов показало, что линейная регрессия обладает высокой интерпретируемостью, но ограничена предположением линейности, KNN чувствителен к масштабированию и шуму, решающее дерево склонно к переобучению без ограничений, тогда как ансамблевые методы, особенно Random Forest, обеспечивают наилучшее соотношение устойчивости и качества.

Анализ гипотез улучшения бейзлайна подтвердил, что препроцессинг данных и инженерия признаков являются ключевыми факторами повышения качества моделей, подбор гиперпараметров особенно важен для регрессионных задач, а оптимизация порога классификации эффективна при дисбалансе классов. Реализация собственных версий алгоритмов показала сопоставимое качество по сравнению с библиотечными реализациями из sklearn, при этом небольшие расхождения в метриках объясняются упрощённой реализацией и отсутствием внутренних оптимизаций.

В целом результаты работы подтверждают, что качество моделей машинного обучения в первую очередь определяется качеством данных, корректной постановкой задачи и выбором метрик, тогда как использование более сложных алгоритмов без улучшенного бейзлайна не гарантирует повышения качества, а собственные реализации обладают высокой образовательной ценностью, позволяя глубже понять внутренние механизмы работы алгоритмов.