ГУАП

КАФЕДРА № 43

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Дожность |  |  |  | В.Ю. [Скобцов](https://pro.guap.ru/inside/profile/11959) |
| канд. техн. наук, доцент |  | подпись, дата |  | инициалы, фамилия |

|  |  |
| --- | --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №1 |  |
| Разведочный и регрессионный анализ данных на основе нейросетевых моделей |  |
| по дисциплине: [Интеллектуальный анализ данных на основе методов машинного обучения](https://pro.guap.ru/inside/students/subjects/3398495) |  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. | 4236 |  |  |  | Л. Мвале |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург

2025

**ЗАДАНИЕ**

Дан многомерный размеченный набор данных. Необходимо выполнить регрессионный анализ данных на основе полносвязной нейросетевой модели и нейросетевой модели, указанной в варианте, в соответствии со следующей последовательностью этапов.

1. Загрузить необходимые пакеты и библиотеки.

2. Загрузить данные из указанного источника.

3. Выполнить разведочный анализ данных в соответствии с этапами описанными в файле Этапы проекта машинного обучения в примерах.pdf:\

a. Ознакомление с данными с помощью методов описательной статистики;\

b. Выполнить визуализацию данных одномерную для понимания распределения данных и многомерную для выяснения зависимостей между признаками;\

c. При необходимости выполнить очистку данных одним из методов.\

d. Проанализировать корреляционную зависимость между признаками;\

e. Поэкспериментировать с комбинациями атрибутов. При необходимости добавить новые атрибуты в набор данных.\

f. Выполнить отбор существенных признаков. Сформировать набор данных из существенных признаков.\

g. При необходимости преобразовать текстовые или категориальные признаки одним из методов.\

h. Выполнить преобразование данных для обоих наборов (исходного и сформированного) одним из методов по варианту.

4. Анализ выполняется для исходного набора данных, преобразованного исходного набора данных, построенного набора данных и преобразованного построенного набора данных. Во всех наборах данных выделить обучающую, проверочную (валидационную) и тестовую выборки данных.

5. Сравнить качество полносвязной нейросетевой регрессионной модели и регрессионной нейросетевой модели, указанной в варианте, на обучающей и валидационной выборках для всех наборов данных, включая их преобразованные варианты. Для оценки качества моделей использовать метрики: корень из среднеквадратичной ошибки, коэффициент детерминации R2.

6. Для лучшей модели на лучшем наборе данных оценить качество на тестовом наборе.

7. Для лучшей модели на лучшем наборе данных выполнить Grid поиск лучших гиперпараметров регрессионной нейросетевой модели на обучающей и валидационной выборках. Определить значения лучших гиперпараметров.

8. Определить показатели качества полученной в результате Grid поиска регрессионной нейросетевой модели на тестовом наборе. Сравнить показатели качества лучшей модели на лучшем наборе данных до поиска гиперпараметров и после поиска гиперпараметров.

9. Сделать выводы по проведенному анализу.

**ПОСТАНОВКА ИНДИВИДУАЛЬНЫЙ ВАРИАНТ ЗАДАНИЯ**

Вариант 5

**Цель проекта**: Провести регрессионный анализ данных телемониторинга болезни Паркинсона для прогнозирования целевого признака `total\_UPDRS` на основе нейросетевых моделей.

**Формальное описание задачи**

Проблема: Построение регрессионной модели для прогнозирования общего показателя UPDRS (Unified Parkinson's Disease Rating Scale) на основе голосовых характеристик пациентов.

Целевая переменная:`total\_UPDRS`

Входные признаки: Все признаки из набора данных, за исключением:

- `motor\_UPDRS` (исключен по условию варианта)

- `index`, `subject#` (служебные идентификаторы)

**Основные этапы решения:**

1. Разведочный анализ данных (EDA) с описательной статистикой и визуализацией

2. Предобработка данных, включая стандартизацию и отбор признаков

3. Построение и сравнение двух типов нейросетевых моделей:

   - Полносвязная нейронная сеть(MLP) Multi-Layer Perceptron

   - Двунаправленная LSTM-сеть

4. Оптимизация гиперпараметров и оценка качества моделей

Мотивация и применение

\*\*Практическая значимость: Разработанная модель может быть использована для:

- Неинвазивного мониторинга прогрессирования болезни Паркинсона

- Объективной оценки эффективности лечения

- Раннего выявления ухудшения состояния пациентов

Метрики качества:

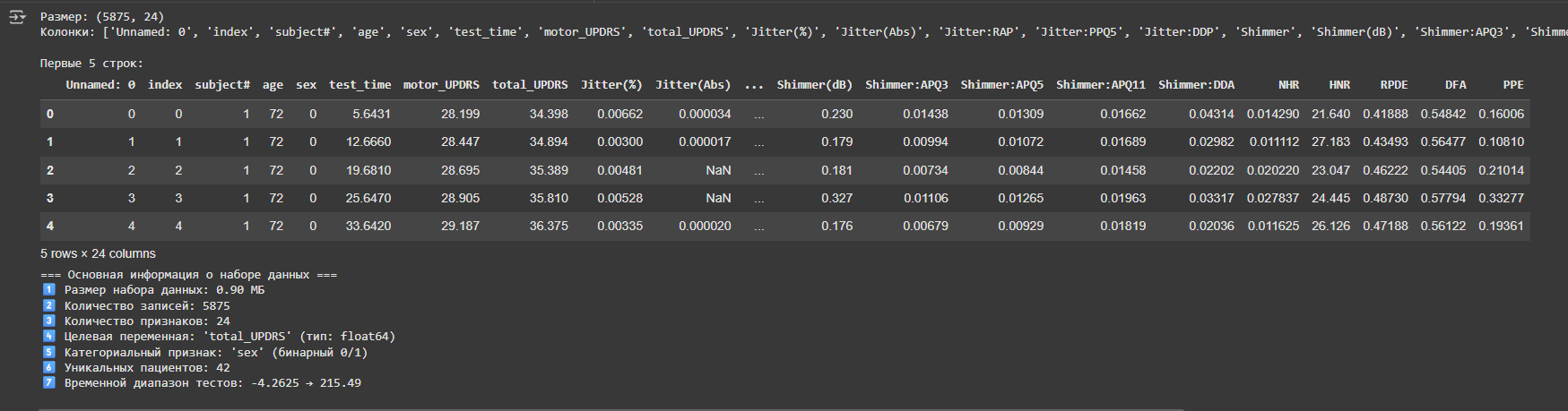
- RMSE (корень из среднеквадратичной ошибки)

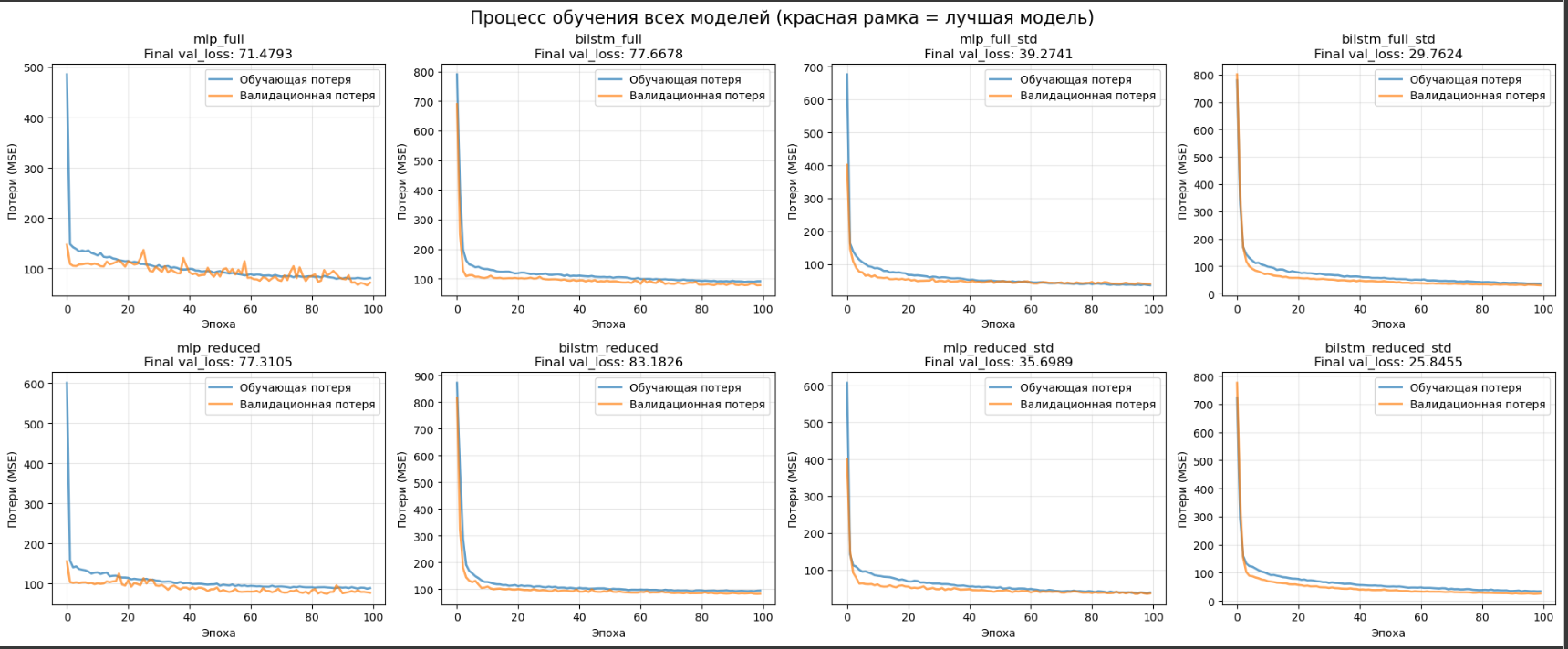
- R² (коэффициент детерминации)

Предметная область

Болезнь Паркинсона - прогрессирующее неврологическое заболевание, где показатель UPDRS является стандартным инструментом оценки тяжести симптомов. Голосовые характеристики (jitter, shimmer, HNR и др.) являются важными маркерами двигательных нарушений, что делает возможным построение регрессионных моделей для прогнозирования общего состояния пациента.

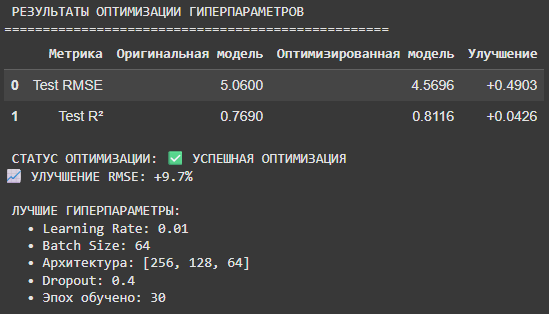
**Примеры графиков, рассчитанных в лабораторной**

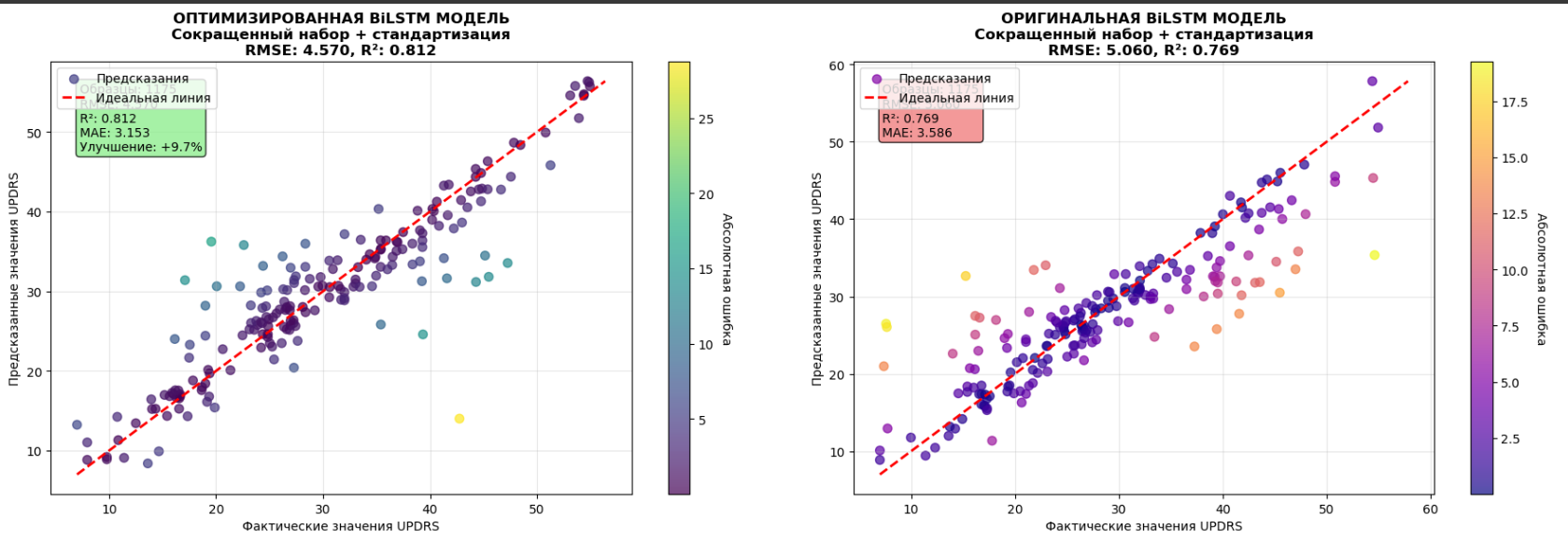




ДЕТАЛЬНЫЕ РЕЗУЛЬТАТЫ ВСЕХ МОДЕЛЕЙ:







**Заключение по регрессионному анализу данных телемониторинга болезни Паркинсона**

Проведенное исследование продемонстрировало высокую эффективность нейросетевых моделей для прогнозирования показателя total\_UPDRS на основе голосовых характеристик. Наилучшие результаты показала оптимизированная двунаправленная LSTM (BiLSTM), обученная на стандартизированном наборе с отобранными признаками.

Ключевые выводы:

- BiLSTM превзошла полносвязную сеть по всем метрикам, особенно на стандартизированных данных

- Оптимальная конфигурация: сокращенный набор признаков + стандартизация

- Гиперпараметрическая оптимизация улучшила RMSE на 9.7% (с 5.0600 до 4.5696)

- Финальное качество: R² = 0.8116, что объясняет 81% дисперсии целевой переменной

Практическая значимость:  
 Разработанная модель позволяет неинвазивно оценивать тяжесть симптомов Паркинсона через анализ голоса. Это открывает возможности для дистанционного мониторинга пациентов, объективной оценки эффективности терапии и раннего выявления прогрессирования заболевания. Модель может интегрироваться в телемедицинские платформы для снижения нагрузки на клиники и улучшения качества жизни пациентов.

Рекомендуется к внедрению оптимизированная BiLSTM как наиболее точное и надежное решение для прогнозирования UPDRS по голосовым маркерам.

Ссылка на лабораторию: https://drive.google.com/file/d/1vvk3juh56WVLWhs0k1JLmCYff-TsVmyE/view?usp=drive\_link