ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

ВЫСШАЯ ИНЖЕНЕРНАЯ ШКОЛА

Выпускная квалификационная работа:

ПРЕДСКАЗАНИЕ ЛИТЕРАТУРНОГО ЖАНРА НА ОСНОВЕ КНИЖНОЙ АННОТАЦИИ

по программе профессиональной переподготовки:

«Анализ данных на языке Python»

Выполнила: Ковылова Полина Григорьевна

Руководитель: Семендяев Родион Юрьевич

Санкт-Петербург 2025

Проблема и актуальность

Современный книжный рынок стремительно развивается благодаря цифровизации. Платформы вроде Goodreads и Amazon ежедневно обрабатывают миллионы книг, и ручная классификация жанров становится неэффективной. Почему это сложно?

- Многожанровость
- Компактность(Неполнота)
- Субъективность
- Шум
- Многоязычность
- Эволюция жанров

Решение — использование методов обработки естественного языка (NLP) и машинного обучения (ML) для мультиметочной классификации, которая позволяет присваивать книге сразу несколько жанров.

Цель и задачи

Цель работы - создание модели машинного обучения на Python для мультиметочной классификации жанров по аннотациям с высокой точностью, пригодной для использования в издательских и рекомендательных системах. Для этого были поставлены следующие задачи:

- Собрать датасет
- Провести предобработку данных
- Разработать и протестировать шесть моделей машинного обучения
- Оценить модели с использованием метрик
- Создать визуализации
- Сформулировать рекомендации

Сбор и предобработка данных

Исследуемый датасет включал 4378 записей с 277 уникальными жанрами, собранных с платформы Goodreads. Этапы предобработки:

- Очистка аннотаций
- Токенизация
- Удаление стоп-слов
- Лемматизация
- Кодирование жанров
- Векторизация текста

```
Loaded 4398 records from Goodreads
After removing missing and short annotations: 4387 records
```

	description	genres
0	It was such a hot summer. The sky was deep blu...	['Picture Books', 'Childrens', 'Fiction', 'Ani...
1	Martin Urban is a quiet bachelor with a comfor...	['Mystery', 'Fiction', 'Crime', 'Thriller', 'B...
2	In Book Two of the Heechee Saga, Robinette Bro...	['Science Fiction', 'Fiction', 'Space Opera', ...
3	A pause in the civil war offers Shrewsbury's t...	['Mystery', 'Historical Fiction', 'Fiction', '...
4	A computerized horse-betting system falls into...	['Mystery', 'Fiction', 'Thriller', 'Crime', 'S...

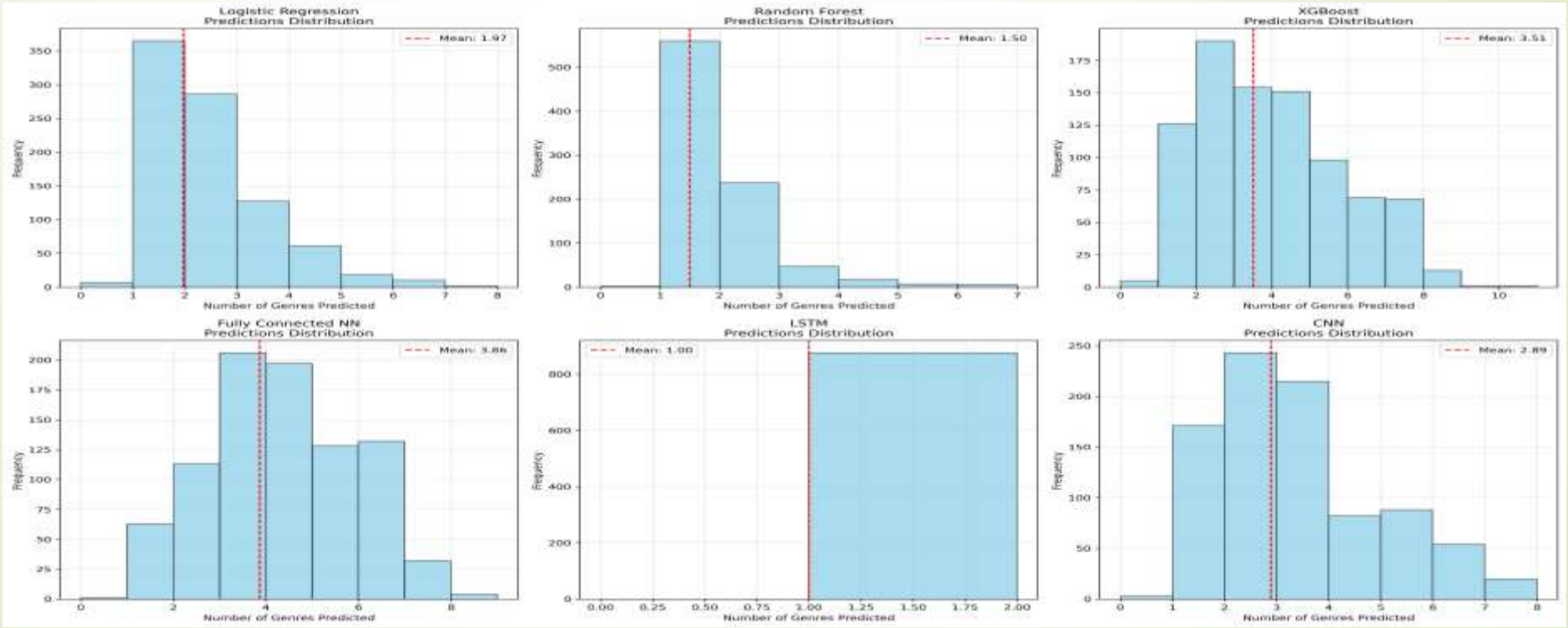
Выбор и обучение моделей

Классические модели

- **Logistic Regression**
- **Random Forest**
- **XGBoost**

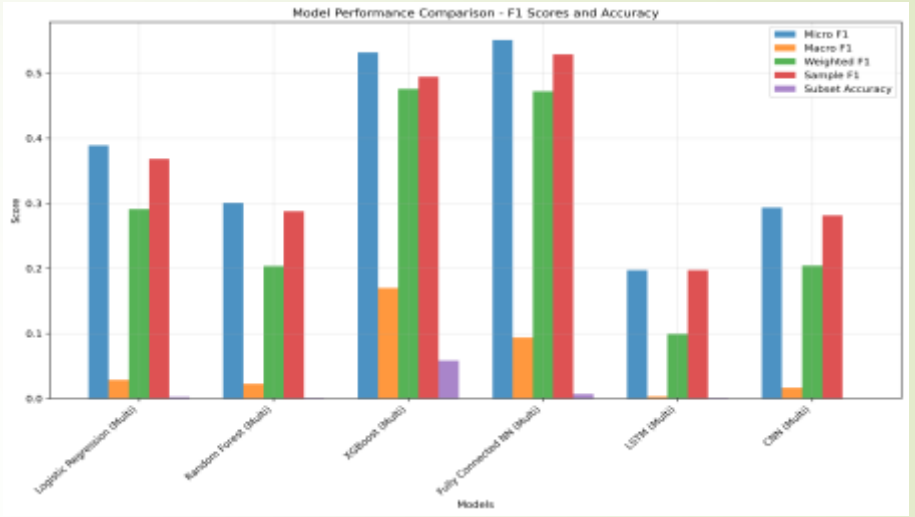
Нейронные сети

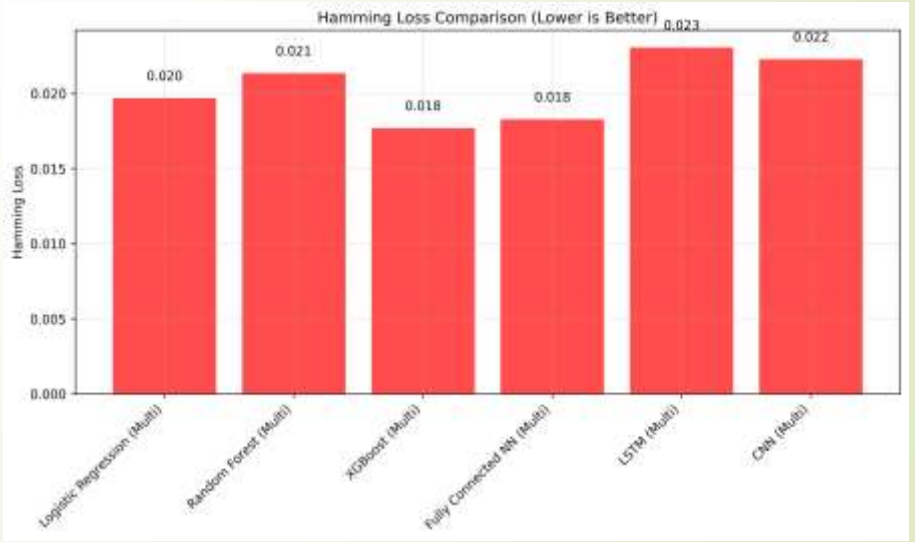
- **Fully Connected Neural Network (Fully Connected NN)**
- **LSTM**
- **CNN**



Оценка метрик

- **Hamming Loss** - доля ошибок, критерий успеха ≤ 0,03
- **Subset Accuracy** - точное совпадение всех жанров
- **Micro F1** – общая оценка точности, наш критерий успеха >0,50 для классических моделей, >0,55 для нейронных
- **Macro F1** - равный вес редким жанрам
- **Weighted F1** - учитывает частоту жанров, давая больше веса популярным
- **Sample F1** - оценивает точность для каждой книги отдельно, а потом усредняет по всем книгам





Анализ результатов

XGBoost : Лучшая классическая модель благодаря градиентному бустингу и взвешиванию классов. Высокая точность частых жанров, умеренное время обучения (~15–25 минут).

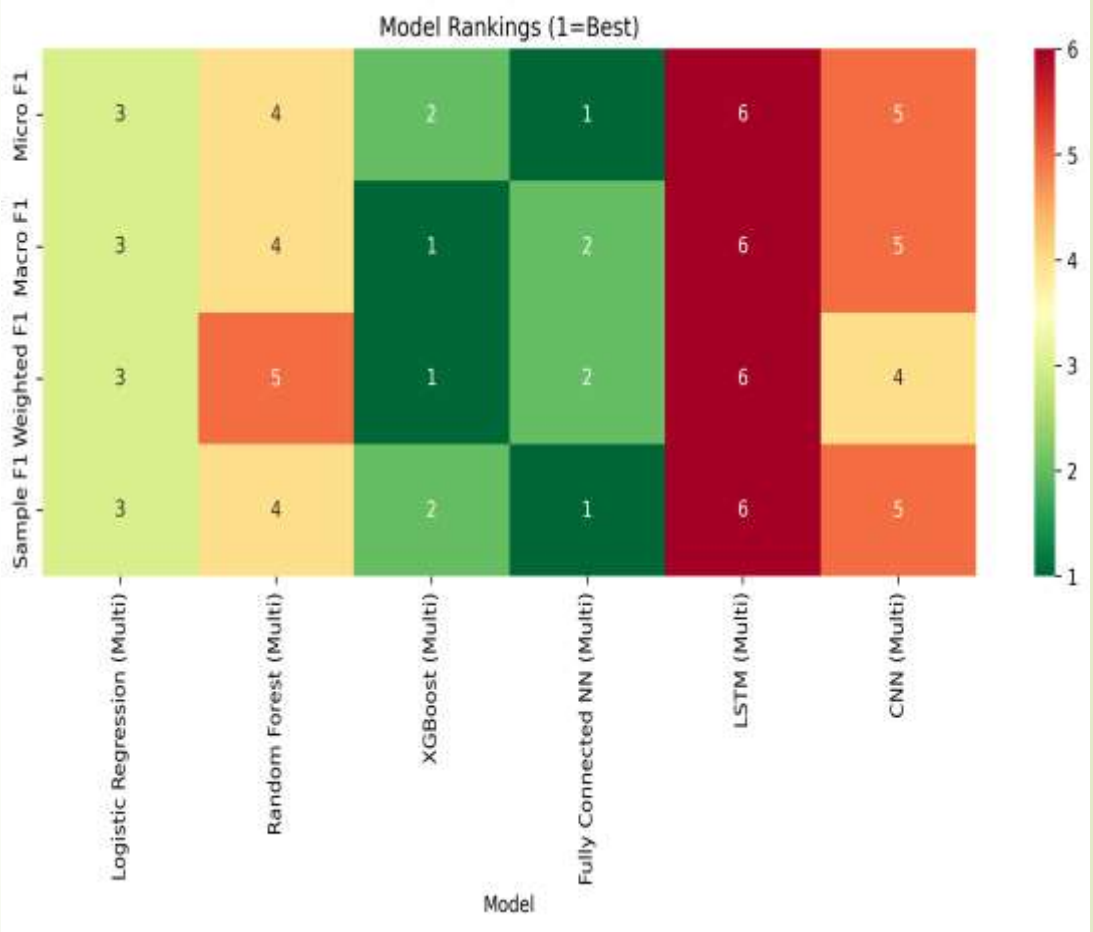
Fully Connected NN: Лидирует среди нейронных моделей, эффективна для частых жанров Быстрое обучение (~3 минуты).

CNN: Отличается средней производительностью, быстрее LSTM (~2 минуты), но хуже улавливает контекст в длинных аннотациях

Logistic Regression: Быстрая (~5–10 минут), но низкая точность из-за линейной природы, эффективна только для явных маркеров (Fiction).

Random Forest: Худшая среди классических моделей, ограничена неспособностью улавливать сложные зависимости.

LSTM: Худшая модель из-за медленной сходимости, чувствительности к гиперпараметрам и ограниченных данных (~18 минут обучения).



Рекомендации

- Увеличение объема данных для улучшения производительности, особенно для редких жанров.
- Использование более продвинутых моделей, таких как BERT, при наличии GPU.
- Оптимизация гиперпараметров XGBoost и Fully Connected NN.
- Интеграция дополнительных данных, таких как отзывы читателей или информация об авторах.
- Тестирование модели на реальных платформах для сбора обратной связи.

Спасибо за внимание!