

ПРОГНОЗИРОВАНИЕ ВЕРОЯТНОСТИ ОТЧИСЛЕНИЯ СТУДЕНТОВ

Третьяков К.С.¹, Девин Г.А.²

¹*Национальный исследовательский Томский политехнический университет, ИШИТР, 8К12,
e-mail: kst12@tpu.ru*

²*Национальный исследовательский Томский политехнический университет, ИШИТР, 8К12,
e-mail: gad7@tpu.ru*

Научный руководитель: Брагин А.Д., старший преподаватель

Аннотация

Работа посвящена проблеме высокого процента отчислений студентов. Основная цель исследования - разработка метода для прогнозирования вероятности отчисления студентов. В исследовании были рассмотрены различные алгоритмы рекуррентной нейронной сети, включая такие модификации, как GRU и LSTM. Оценка производительности моделей проводилась с помощью ROC-кривых, accuracy, recall, f1_score. Это исследование может помочь учебным заведениям в предотвращении отчислений, опираясь на прогностические данные.

Ключевые слова: рекуррентная нейронная сеть (RNN), GRU (Gated Recurrent Unit), LSTM (Long Short-Term Memory), прогнозирование отчислений студентов, кросс-валидация и оценка производительности.

Введение

Проблема высокого процента отчислений студентов ставит перед учебными заведениями серьезные вызовы. Для решения этой проблемы, прогнозирование вероятности отчисления студентов становится важным инструментом, позволяющим выявлять рискованные ситуации, предпринимать меры по улучшению успеваемости и сохранению студентов, а также эффективно распределять ресурсы и создавать программы поддержки.

Основной целью исследования является создание модели, способной предсказывать вероятность отчисления студентов на основе доступных данных об их активности и успеваемости в учебном процессе. После прогнозирования вероятности отчисления модель будет использоваться для выявления студентов с высоким риском отчисления, что позволит учебным заведениям проводить более эффективную работу с такими студентами.

Описание алгоритма

Для прогнозирования отчислений студентов были использованы наборы данных, содержащие информацию о различных аспектах студенческой активности. Эти наборы данных представляли собой совокупность нескольких наборов данных, включающих следующую информацию: студент, дата рождения, пол, группа, дисциплина, семестр, учебный год, оценка, специальность, форма обучения, пропуск занятий.

Для того чтобы данные стали пригодными для анализа и моделирования, мы провели ряд этапов предобработки данных. Эти этапы включали в себя очистку данных от выбросов и отсутствующих значений, нормализацию числовых признаков, кодирование категориальных признаков и создание дополнительных признаков.

Проведенные этапы помогли улучшить качество данных, заполнить пропущенные значения оценок, исключить ненужные записи, а также добавить контекст для более полного анализа.

Для создания модели прогнозирования вероятности отчисления студентов, мы рассмотрели несколько модификаций рекуррентно нейронной сети: GRU, LSTM и базовая модель RNN.

RNN:

Рекуррентные нейронные сети (RNN) — это класс нейронных сетей, специально разработанный для работы с последовательными данными, такими как тексты, временные ряды или последовательности действий. Они отлично подходят для анализа данных, где важен контекст и зависимость от предыдущих шагов.

Основная идея рекуррентных нейронных сетей заключается в использовании обратных связей, позволяющих передавать информацию о предыдущих состояниях сети на последующие шаги. Это делает их способными к анализу и обработке последовательных данных произвольной длины (рис. 1).

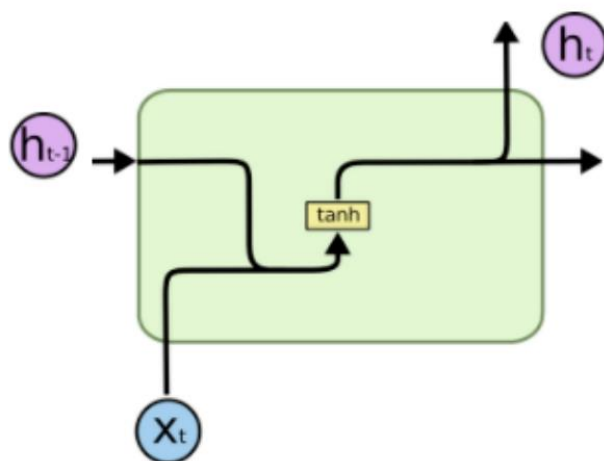


Рис. 1. Базовая архитектура RNN

Преимущества использования рекуррентных нейронных сетей в вашей работе о прогнозировании отчислений студентов:

1. Учет последовательности данных: Использование RNN позволяет учитывать последовательность академических и иных данных о студентах, таких как оценки за прошлые курсы. Это важно, поскольку многие факторы могут влиять на решение студента об отчислении, и они могут быть взаимосвязаны и изменяться с течением времени.

2. Обработка переменной длины входных данных: RNN легко справляются с варьирующимися по длине входными последовательностями данных, что важно для анализа данных о студентах, поскольку у них может быть разное количество семестров, в течение которых они присутствовали в университете, разное количество предметов и т.д.

Также мы решили проанализировать работу модификаций RNN, таких как GRU и LSTM. GRU (Gated Recurrent Unit) и LSTM (Long Short-Term Memory) — это две основные модификации базовой архитектуры RNN (рис. 2.)

LSTM: LSTM является более сложной модификацией RNN, которая включает в себя специальные "ворота" для контроля потока информации в сети. Они способны более эффективно управлять долгосрочными зависимостями и предотвращать проблемы с затуханием и взрывом градиента.

GRU: GRU представляет собой упрощенную версию LSTM, которая объединяет в себе некоторые из ее механизмов, но с меньшим количеством параметров. Они имеют механизмы, которые помогают предотвратить затухание градиента и сохранить информацию на протяжении времени. Анализ работоспособности этих модификаций позволил нам выбрать наиболее подходящий вариант для нашей задачи прогнозирования отчислений студентов.

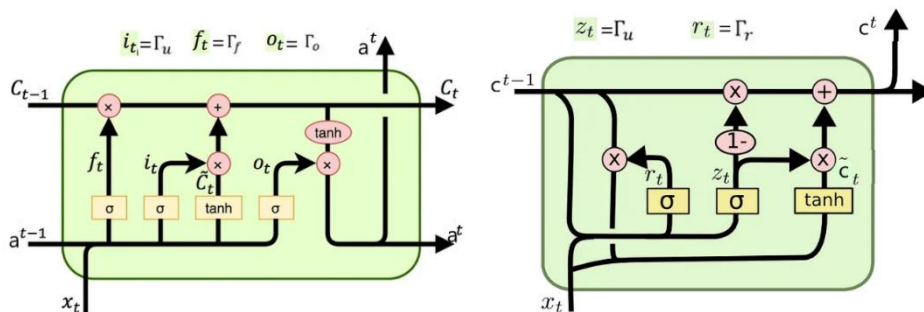


Рис. 2. Архитектура LSTM и GRU

Для анализа наших данных мы применили метод главных компонент (PCA). Этот метод позволяет снизить размерность данных, сохраняя при этом основные характеристики и различия между наблюдениями. Мы использовали PCA для уменьшения размерности нашего набора данных до одной

главной компоненты. Однако, после применения PCA мы решили отказаться от этого подхода, так как обнаружили, что без использования PCA наши модели показывают более высокие показатели точности. Также использовали стандарт скаллер для масштабирования признаков.

Базовая модель, которую мы использовали в нашей работе, представляла собой рекуррентную нейронную сеть (RNN) с функцией активации гиперболического тангенса. Для предотвращения переобучения мы использовали слой дропаута с коэффициентом 0.1. Для получения вероятности класса мы использовали функцию активации сигмоиду. В качестве метрики для оценки модели мы применяли бинарную кросс-энтропию, а в качестве оптимизатора использовали Adam. Результаты базовой модели приведены в таблице 1.

После тестирования первой модели мы решили изменить структуру нашей модели, заменив базовую RNN на LSTM. Результаты модели LSTM приведены в таблице 1.

Также провели тестирование с моделью GRU, результат тестирования приведены в таблице 1.

Таблица 1

Результаты 1 тестирования

	Accuracy	Precision	Recall	F1-score	AUC-ROC
Basic RNN	78 %	89 %	70 %	77 %	81 %
GRU	80 %	93 %	68 %	79 %	83 %
LSTM	83 %	93 %	74 %	83 %	84 %

При проведении тестирования различных моделей для анализа данных было обнаружено, что LSTM продемонстрировала лучшие результаты. В нашем контексте, где требуется анализировать данные с длинными временными зависимостями или сложными шаблонами, LSTM представляет собой привлекательную модель. Её способность сохранять информацию на протяжении длительных интервалов времени позволяет ей эффективно улавливать и использовать контекстуальные зависимости между данными.

После первого этапа тестирования мы внесли изменения в итоговый набор данных, добавив информацию о пропусках и времени, отведенном на дисциплину. Мы также провели стандартизацию данных и убрали уменьшение размерности PCA. Результаты тестирования приведены в таблице 2.

Таблица 2

Результаты 2 тестирования

	Accuracy	Precision	Recall	F1-score	AUC-ROC
Basic RNN	83 %	95 %	72 %	82 %	83 %
GRU	83 %	93 %	73 %	82 %	84 %
LSTM	83 %	95 %	73 %	82 %	84 %

После завершения второго этапа тестирования мы расширили итоговый объем данных, включив дополнительную информацию о типе финансирования студента (бюджетная или договорная основа), возрасте студента на момент поступления, а также информацию о том, являлся ли студент вольным слушателем. До третьего этапа тестирования наш подход к оценке успеха студентов ограничивался анализом их результатов только в последнем семестре. Однако в ходе третьего этапа тестирования мы внесли изменения, начав оценивать метрики по отдельным семестрам. Результаты представлены в Таблице 3.

Результаты 3 тестирования

	Accuracy	Precision	Recall	F1-score	AUC-ROC
1 семестр					
Basic RNN	65 %	62 %	92 %	74 %	72 %
GRU	63 %	61 %	92 %	73 %	70 %
LSTM	69 %	69 %	82 %	75 %	72 %
2 семестр					
Basic RNN	68 %	64 %	84 %	73 %	75 %
GRU	69 %	65 %	84 %	73 %	74 %
LSTM	73 %	72 %	78 %	75 %	76 %
3 семестр					
Basic RNN	73 %	71 %	63 %	67 %	72 %
GRU	72 %	66 %	74 %	69 %	73 %
LSTM	73 %	69 %	69 %	69 %	73 %
4 семестр					
Basic RNN	73 %	69 %	53 %	60 %	69 %
GRU	71 %	61 %	61 %	61 %	72 %
LSTM	72 %	63 %	58 %	60 %	71 %
5 семестр					
Basic RNN	74 %	67 %	45 %	54 %	67 %
GRU	74 %	63 %	55 %	59 %	69 %
LSTM	74 %	65 %	50 %	57 %	69 %
6 семестр					
Basic RNN	75 %	67 %	37 %	47 %	64 %
GRU	74 %	59 %	43 %	50 %	67 %
LSTM	74 %	61 %	39 %	46 %	64 %
7 семестр					
Basic RNN	77 %	63 %	33 %	43 %	63 %
GRU	75 %	54 %	39 %	45 %	63 %
LSTM	76 %	57 %	37 %	46 %	64 %
8 семестр					
Basic RNN	77 %	57 %	28 %	37 %	60 %
GRU	75 %	50 %	35 %	41 %	60 %
LSTM	76 %	53 %	31 %	39 %	61 %

Из результатов третьего этапа тестирования моделей LSTM, GRU и Basic RNN для анализа успеваемости студентов по семестрам следует, что модель LSTM в целом демонстрирует наилучшую эффективность по сравнению с GRU и Basic RNN. Начальные семестры характеризуются более количеством учащихся, так как еще не произошло отчисление студентов, а к 4 курсу количество студентов, которые собираются отчислиться, уменьшается. Это оказывало влияние на точность моделей в предсказании успеваемости студентов на более поздних этапах учебы.

Заключение

В заключение, использование рекуррентных нейронных сетей (RNN) для анализа данных об отчислениях студентов обосновано исходя из их способности учитывать временные зависимости в данных, обрабатывать последовательные и переменной длины входные данные, а также за счет механизмов запоминания долгосрочных зависимостей. Эти модели позволяют эффективно

моделировать динамику студенческой активности и делать прогнозы относительно вероятности отчисления студентов, что может быть важным для принятия решений в образовательном учреждении. Кроме того, важно учитывать метрики производительности моделей, такие как ROC-кривые, accuracy, recall, и f1_score, для оценки качества модели и сравнения её с другими алгоритмами.

Список использованных источников

1. Understanding LSTM Networks // colah's blog URL: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (дата обращения: 27.03.2024).
2. Understanding Bidirectional RNN in PyTorch // towardsdatascience.com URL: towardsdatascience.com/understanding-bidirectional-rnn-in-pytorch-5bd25a5dd66 (дата обращения: 26.03.2024).
3. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности: справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков. М.: Финансы и статистика. – 1989. – 607 с.
4. Воронцов К.В. Комбинаторные оценки качества обучения по прецедентам // Докл. РАН. – 2004. – Т. 394. – № 2. – С. 175–178.
5. Рекуррентная нейронная сеть (RNN): виды, обучение, примеры // neurohive.io URL: <https://neurohive.io/ru/osnovy-data-science/rekurrentnye-nejronnye-seti/> (дата обращения: 26.03.2024).
6. RNN vs GRU vs LSTM // Medium URL: <https://medium.com/analytics-vidhya/rnn-vs-gru-vs-lstm-863b0b7b1573> (дата обращения: 20.03.2024).