

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

**«Дальневосточный федеральный университет»**

(ДВФУ)

**ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ (ШКОЛА)**

**Департамент математического и компьютерного моделирования**

Анализ вовлеченности в онлайн-курсы через поведенческие паттерны

**КУРСОВАЯ РАБОТА**

по образовательной программе подготовки бакалавров

по направлению 01.03.02 «Прикладная математика и информатика»

профиль «Системное программирование»

|  |  |
| --- | --- |
| Работа защищена  с оценкой             \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_            Регистрационный номер \_\_\_\_\_\_\_  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025г. | Студент группы     № Б9122-01.03.02сп  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Носков Я.В.  (подпись)                                               (ФИО)  «\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2025г.      Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   (должность,  ученое звание)  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Пак Т. В.\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (подпись)                                               (ФИО)  «\_\_\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025г. |

г. Владивосток

2025

Оглавление

[Введение 2](#_Toc198503020)

[1 Подготовка и анализ датасета. 2](#_Toc198503021)

[1.1 Структура сырых данных 2](#_Toc198503022)

[1.2 Подготовка данных 2](#_Toc198503023)

[1.3 Визуализация и анализ данных 2](#_Toc198503024)

[2 Поиск частых последовательностей. PrefixSpan 2](#_Toc198503025)

[2.1 Sequence mining 2](#_Toc198503026)

[2.2 Алгоритм PrefixSpan 2](#_Toc198503027)

[2.3 Результаты 2](#_Toc198503028)

[3 Классификация поведения студентов. 2](#_Toc198503029)

[3.1 Постановка задачи 2](#_Toc198503030)

[3.2 Подготовка данных 2](#_Toc198503031)

[3.3 Random Forest Classifier 2](#_Toc198503032)

[3.4 Gradient Boosting 2](#_Toc198503033)

[3.5 Интерпретация модели с помощью SHAP 2](#_Toc198503034)

[Заключение 2](#_Toc198503035)

[Список литературы 2](#_Toc198503036)

[Приложение 2](#_Toc198503037)

# Введение

В последние годы массовое онлайн-обучение стало неотъемлемой частью современного образования. Благодаря платформам MOOCs (Massive Open Online Courses) миллионы студентов по всему миру получают доступ к курсам ведущих университетов и компаний, что делает образование более доступным и гибким. Однако, несмотря на растущую популярность, онлайн-обучение сталкивается с серьезной проблемой — высокой долей отсева. Исследования показывают, что значительная часть слушателей прекращает обучение, не завершив курс, что снижает эффективность таких образовательных форматов. В условиях отсутствия прямого контакта с преподавателями и группой, ключевым источником информации о студенте становятся цифровые следы. В связи с этим возникает необходимость в разработке методов анализа вовлеченности студентов и прогнозирования отсева на основе их цифровых следов — данных о взаимодействии с платформой.

Объектом исследования выступают поведенческие данные (логи) студентов онлайн-курсов, включающие их действия, время действия.

Предмет исследования — MOOC dataset китайской образовательной платформы XuetangX в периоде от 2015-2017 годов.

Цель исследования — анализ вовлеченности студентов в онлайн-курсах через их поведенческие паттерны и выявление различий между теми, кто завершил обучение, и теми, кто его бросил.

Для достижения этой цели были поставлены следующие задачи:

1. Провести предварительный анализ датасета MOOCs Learning Analytics, выявить различия в активности между студентами, завершившими курс, и теми, кто его бросил.

2. Применить методы sequence mining (алгоритм PrefixSpan) для выделения частых паттернов поведения и определения ключевых точек отсева.

3. Построить агрегированные признаки (метрики вовлеченности) на уровне студент-курс, включая количество действий, длительность сессий, тип курса и профиль студента.

4. Разработать модель классификации, прогнозирующую вероятность отсева на основе поведенческих данных.

5. Проанализировать полученные результаты и предложить рекомендации по оптимизации структуры курсов и повышению вовлеченности студентов.

Методы исследования включают:

• Sequence mining (алгоритм PrefixSpan) для анализа последовательностей действий студентов и выявления типичных сценариев взаимодействия с курсом.

• Анализ агрегированных признаков, включающий расчет метрик вовлеченности (активность, длительность сессий, частота взаимодействия с различными типами контента).

• Машинное обучение (классификация) для прогнозирования отсева на основе поведенческих данных.

• Визуализация данных для интерпретации паттернов вовлеченности и критических точек отсева.

Основные сложности, с которыми столкнулось исследование:

• Обработка больших объемов логов и оптимизация представления данных.

• Корректное определение метрик вовлеченности, отражающих реальные различия в поведении студентов.

Практическая значимость работы заключается в возможности создания системы раннего предупреждения, которая будет автоматически выявлять студентов из группы риска и предлагать персонализированные вмешательства. Это особенно важно для масштабных образовательных платформ, где ручной мониторинг прогресса каждого обучающегося невозможен.

Таким образом, работа сочетает в себе описательную аналитику (выявление паттернов) и прогнозное моделирование, что позволяет не только понять причины отсева, но и предложить инструменты для его снижения.

# 1 Подготовка и анализ датасета.

Для анализа вовлечённости в онлайн-обучении использовался **MOOC Dataset**, доступный на платформе Kaggle. Он содержит подробную информацию о действиях студентов на онлайн-курсах китайского университета: просмотры видео, навигация по контенту, выполнение заданий и другое. Основной интерес представляют **поведенческие логи**, записанные в виде действия, совершенного в момент времени.

## 1.1 Структура сырых данных

**Таблица логов**

* enroll\_id – уникальный идентификатор прохождения курса пользователем;
* username – идентификатор пользователя;
* course\_id – идентификатор курса;
* session\_id – идентификатор сессии;
* action – тип действия;
* time – время действия;
* truth – бинарный целевой признак (0 — завершил курс, 1 — бросил).

**Таблица пользователей**

* user\_id – идентификатор пользователя;
* gender – пол пользователя
* education – уровень образования
* birth – дата рождения

**Таблица курсов**

* course\_id – идентификатор курса;
* start – начало курса
* end – конец курса
* course\_type – тип курса (с ментором, без ментора)
* category – категория курса (физика, биология)

## 1.2 Подготовка данных

**Таблица пользователей**

Из-за большого количества пропусков признаки из таблицы пользователей были заменены на признак показывающий заполненность профиля (1 – профиль заполнен, 0 – в профиле есть пропуски).

**Таблица курсов**

Из-за большого количества пропусков признак category был удален. Добавлен новый признак продолжительность курса. Рассчитывающийся по формуле:

|  |  |
| --- | --- |
|  | (1) |

После всех преобразований, датасет для анализа содержит следующие данные:

* enroll\_id
* количество каждого действия отдельно
* суммарное время всех сессий на курс
* законченность профиля
* тип курс
* длительность курса
* truth (метка 0 – курс сдан, 1 – курс брошен)

## 1.3 Визуализация и анализ данных

Для анализа данных был использован язык программирования python и библиотека визуализации matplotlib. Были визуализирован каждый признак на трех графиках. Два графика показывают гистограммы признака при разной метке truth. Третий график показывают разницу между первыми двумя графиками.

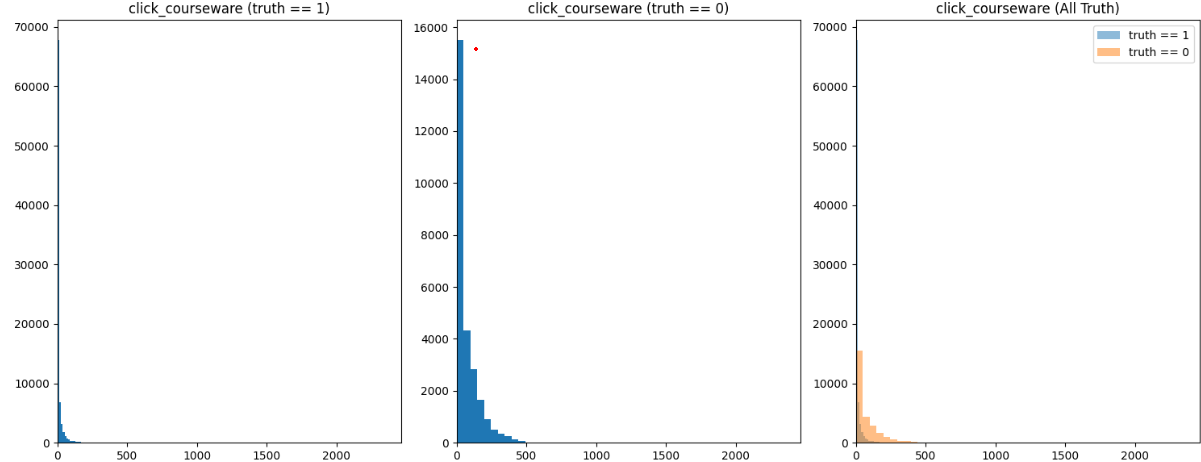


Рисунок 1 – пример визуализации признака из датасета

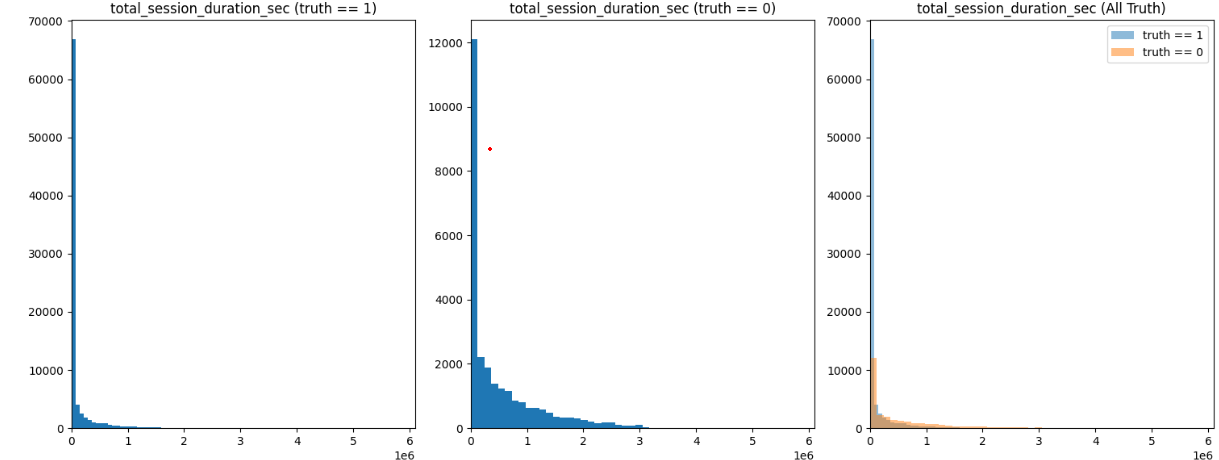


Рисунок 2 – визуализация признака общее время сессий в секундах.

В результате визуализации были удалены признаки: “close\_forum”, “create\_thread”, “delete\_thread” так как они были малоинформативными. По остальным признакам стало очевидно, что существует разница в активности между студентами, которые закончили и бросили курс.

# 2 Поиск частых последовательностей. PrefixSpan

## 2.1 Sequence mining

Sequence pattern mining (поиск частых последовательностей) — это направление анализа данных, направленное на выявление закономерностей в упорядоченных наборах действий или событий. В контексте онлайн-обучения данный подход позволяет анализировать поведение студентов на основе последовательностей их взаимодействий с образовательной платформой. Например, можно обнаружить, что пользователи, которые чаще начинают с просмотра видео, затем делают паузы и переходят к выполнению заданий, с большей вероятностью завершают курс. Напротив, другие паттерны могут коррелировать с высокой вероятностью отсева.

В данной работе sequence mining используется как инструмент для выявления различий в поведенческих паттернах между теми, кто завершил курс, и теми, кто его бросил. Это позволяет не только лучше понять характер поведения учащихся, но и в перспективе внедрить систему раннего оповещения об угрозе отсева.

Формально, последовательность – это упорядоченный набор событий:

*,*

где  **–** элементарное действие пользователя.

В данной работе sequence mining используется как инструмент для выявления различий в поведенческих паттернах между теми, кто завершил курс, и теми, кто его бросил. Это позволяет не только лучше понять характер поведения учащихся, но и в перспективе внедрить систему раннего оповещения об угрозе отсева.

Цель sequence mining **–** найти все подпоследовательности, которые:

* часто встречаются среди всех пользователей;
* могут служить предикторами исхода (например, отсева).

## 2.2 Алгоритм PrefixSpan

В данной работе для анализа поведенческих последовательностей применялся алгоритм **PrefixSpan (Prefix-projected Sequential Pattern Mining)**, один из наиболее эффективных алгоритмов частотного поиска последовательностей.

**Идея алгоритма**

Вместо генерации и проверки всех возможных последовательностей (как в Apriori), PrefixSpan использует идею префиксного проецирования: если последовательность α является префиксом некоторой частой последовательности, то продолжения этой последовательности можно искать только в проекциях баз данных, содержащих α.

**Формальное описание алгоритма**

Пусть:

* — множество возможных действий (алфавит, itemset);
* } – множество всех пользовательских последовательностей;
* *-* последовательность, которая содержится в одной или нескольких последовательностях из *D;*
* support(P) – количество последовательностей в *D*, содержащих .

|  |  |
| --- | --- |
|  | (2) |

Тогда алгоритм работает следующим образом:

1. Инициализация: найти все часто встречающиеся элементы, формирующие шаблоны длины 1:

|  |  |
| --- | --- |
|  | (3) |

1. **Рекурсивное расширение**: для каждого префикса , строится его проецированная база последовательностей:

|  |  |
| --- | --- |
|  | (4) |

где – оставшаяся часть последовательности после вхождения префикса в .

1. На основе находятся допустимые расширения префикса :

* Внутри itemset’а:
* Расширение новым itemset’ом:

1. Для каждого расширения , если , то:

* Добавить в множество частых шаблонов
* Рекурсивно вызвать PrefixSpan с префиксом и

Алгоритм завершается, когда, не остаётся допустимых расширений префикса, удовлетворяющих порогу поддержки .

**Формирование датасета**

Для применения алгоритма PrefixSpan необходимо было подготовить датасет в формате, удобном для поиска последовательностей, то есть необходимо сделать последовательности из действий.

Данные о действиях были сгруппированы по сессиям пользователя и разделены специальными токенами <start> и <end>. После этого действия в сессиях были собраны в последовательности по enroll\_id.



Рисунок 1 пример базы данных с последовательностями для PrefixSpan.

## 2.3 Результаты

После применения алгоритма PrefixSpan с параметрами:

minlen = 1, maxlen = 6 – искались шаблоны длинны 1 – 6 дейтсвий;

frequent = 0.05 – порог частоты последовательностей не менее, чем у 5% студентов.

Были получены часто встречаемые последовательности у тех, кто сдал курс (truth = 0) и не сдал курс (truth = 1).

В результате анализа выяснилось, что распределение последовательностей между двумя группами оказались схожими. Большинство наиболее частых паттернов встречались у обоих групп с примерно одинаковой частотой.

Это говорит о том, что одного только порядка действий недостаточно, чтобы судить о финальном результате. Поведение студентов в рамках одной или нескольких сессий может быть схожим.

Таким образом, поведенческие паттерны не дали ярко выраженных различий. Это важно, так как результат говорит о том, что необходимы дополнительные факторы для построения более точной модели.

# 3 Классификация поведения студентов.

После результатов, полученных с помощью sequence mining, встала задача предсказания вероятности завершения курса студентом на основе действий или интенсивности действий студента.

## 3.1 Постановка задачи

Формально, задача ставится как бинарная классификация:

где x – вектор признаков

Ценность модели – определить “рисковых” слушателей, а также выявить какие действия больше всего влияют на результат модели.

## 3.2 Подготовка данных

Для задачи бинарной классификации было сформировано 2 датасета с агрегированной статистикой. Столбцами датасета являются количество действий. Так же были добавлены дополнительные признаки: тип курса (курс с ментором или без ментора), общая длительность сессий, наличие заполного профиля.

Первый датасет, на основе действий совершенных студентом на протяжении всего курса, необходим, чтобы выявить значимые различия между студентов разных групп.

Второй датасет, на основе действий совершенных в рамках сессий, необходим, чтобы подтвердить то, что действия студентов из разных категорий, совершенных в одной сессии, имеют мало различий.

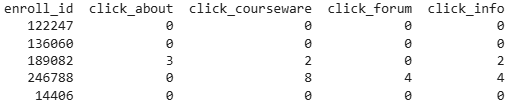


Рисунок 3 – иллюстративный пример части первого датасета.

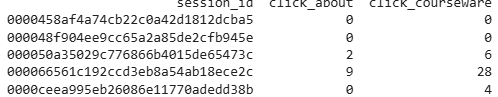


Рисунок 2 - иллюстративный пример части второго датасета.

После этого данные были разделены тренировочные и тестовые в соотношении 70/30.

Для решения задачи были выбраны два ансамблевых метода машинного обучения: Random Forest Classifier, Gradient Boosting.

## 3.3 Random Forest Classifier

Random Forest Classifier - ансамблевый метод, основанный на построении набора решающих деревьев (decision trees) и объединении их предсказаний. Это один из наиболее устойчивых к переобучению алгоритмов классификации, особенно эффективный при работе с табличными данными.

**Алгоритм**

Пусть задана обучающая выборка:

|  |  |
| --- | --- |
|  | (5) |

Модель случайного леса строит *M* независимых деревьев решений , каждое из которых обучается на случайной бутстрэп-выборке из исходных данных. При построении каждого дерева:

* На каждом узле случайно выбирается подмножество признаков ,
* Выбирается наилучшее разбиение по одному из этих признаков,
* Дерево растет до максимальной глубины или до достижения других критериев остановки.

Финальное предсказание модели производится голосованием:

|  |  |
| --- | --- |
|  | (5) |

где:

– отдельное дерево, обученное на бутстрэп-подвыборке;

mode – наиболее часто встречающееся значение.

Достоинства модели:

* Устойчива к переобучению;
* Возможность оценки важности признаков.

**Результаты на первом датасете**

Модель Random Forest показала следующие результаты на тестовой выборке:

* Accuracy: 0.840
* Precision: 0.80
* Recall: 0.74
* F1-score: 0.76
* ROC AUC: 0.74

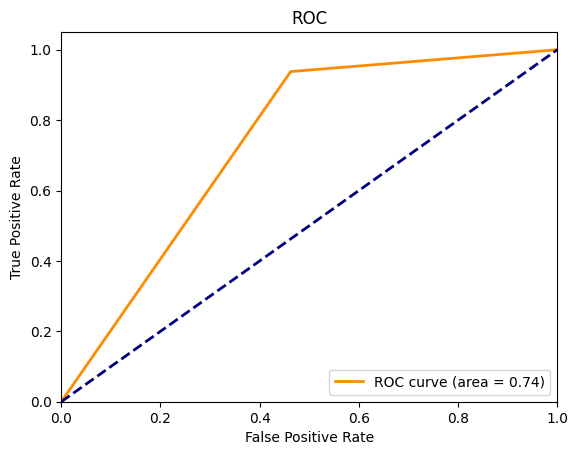


Рисунок 4 – ROC кривая модели Random Forest

**Результаты на втором датасете**

* Accuracy: 0.66
* Precision: 0.6
* Recall: 0.6
* F1-score: 0.6
* ROC AUC: 0.6

Данные результаты показывают, что модель плохо способна определять разницу между теми, кто завершил курс и теми, кто его бросил на основе действий в одной сессии. Это подтверждает, что нет разница между студентами, завершившими и бросившими курс внутри сессий.

После данной проверки, можно сделать вывод, что **использование второго датасета не целесообразно** для достижения хороших показателей метрик.

## 3.4 Gradient Boosting

Gradient Boosting – это метод ансамблирования, при котором деревья решений обучаются **последовательно**, и каждое последующее дерево направлено на исправление ошибок предыдущих.

**Алгоритм**:

На каждой итерации строится модель , которая аппроксимирует отрицательный градиент функции потерь по предыдущему ансамблю. Итеративно улучшается приближение:

|  |  |
| --- | --- |
|  | (6) |
|  | (7) |

где:

* – функция потерь;
* – новое дерево, аппроксимирующее градиент функции потерь;
* – шаг обучения;
* – ансамбль из *m* моделей.

Каждое новое дерево обучается на остатках (градиентах), соответствующих текущим ошибкам модели:

|  |  |
| --- | --- |
|  | (8) |

**Предсказание**:

Окончательное предсказание после *M* итераций:

|  |  |
| --- | --- |
|  | (5) |

**Особенности**

* Позволяет точно учитывать сложные зависимости между признаками;
* Требует тщательной настройки параметров (глубина деревьев, learning rate, число итераций);
* Чувствителен к переобучению при слишком большом числе итераций или отсутствии регуляризации.

**Результаты на первом датасете**

Модель Gradient Boosting показала следующие результаты на тестовой выборке:

* Accuracy: 0.843
* Precision: 0.81
* Recall: 0.74
* F1-score: 0.76
* ROC AUC: 0.74

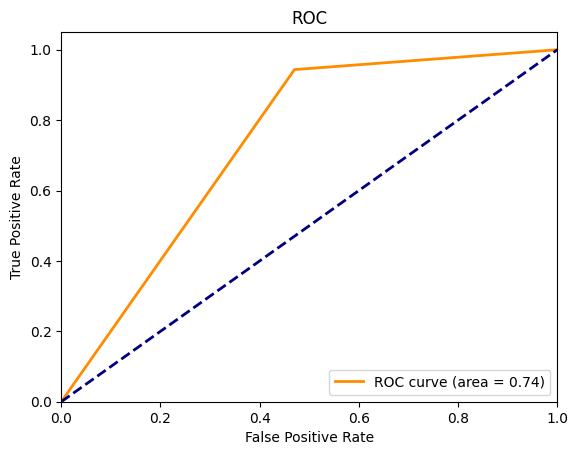


Рисунок 4 – ROC кривая модели Gradient Boosting.

В результате **Random Forest** и **градиентный бустинг** в задачи предсказания отсева студентов продемонстрировали высокую точность классификации, при этом модель Random Forest показала **accuracy = 0.84**, что свидетельствует о хорошей способности различать между теми студентами, которые завершат курс, и теми, кто с высокой вероятностью его покинет. Так как Random Forest более устойчив к выбросам, выберем его.

## 3.5 Интерпретация модели с помощью SHAP

Для более глубокого анализа полученной модели классификации был применён метод SHAP (SHapley Additive ExPlanations), основанный на теории игр. Он позволяет определить вклад каждого признака в предсказание модели для каждого конкретного наблюдения.

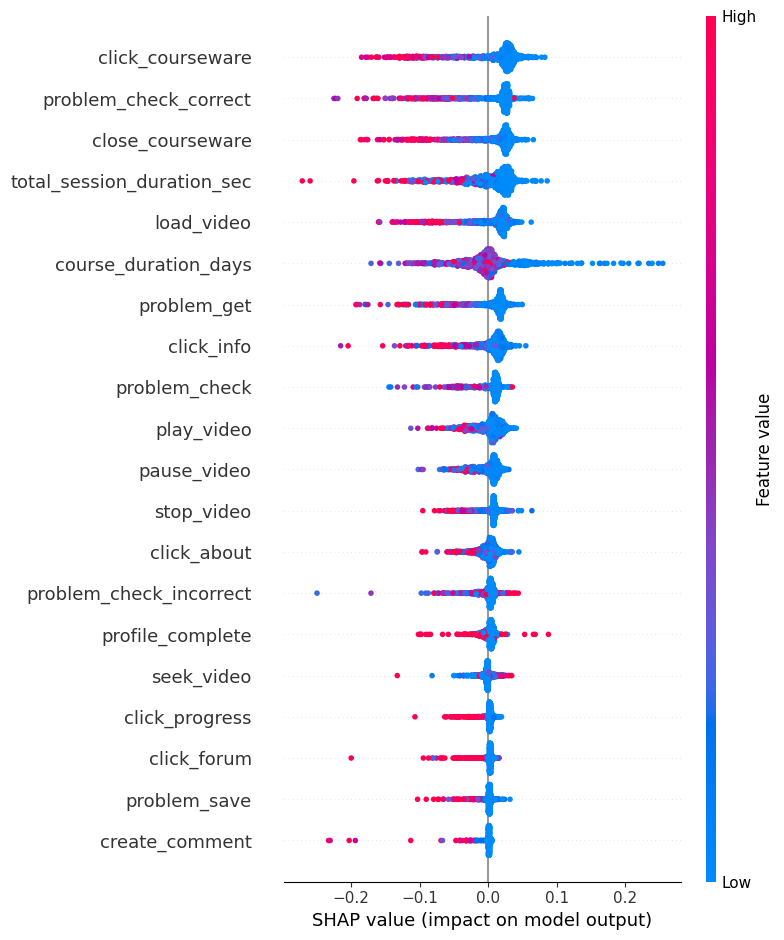


Рисунок 5 – SHAP summary plot важности признаков для модели Random Forest

На графике важности признаков, рассчитанном с помощью shap.summary\_plot, самые важные признаки находятся сверху. При этом цвет означает интенсивность, синий – действие совершено маленькое количество раз, красный – наоборот, большое количество раз. SHAP value показывает, в какую сторону влияет признак при truth = 1.

Согласно графику, чем меньше активность почти каждого действия, тем больше вероятность отнести входящий вектор к классу 1 (не окончил курс).

# Заключение

В данной курсовой работе была рассмотрена задача анализа вовлечённости студентов в онлайн-обучении с использованием методов машинного обучения. На датасета MOOC Dataset с платформы Kaggle были исследованы поведенческие паттерны пользователей и построена модель прогнозирования вероятности завершения курса.

Анализ вовлечённости осуществлялся по двум направлениям. На первом этапе применялись методы **sequence mining**, в частности, алгоритм PrefixSpan, позволяющий выявлять часто встречающиеся последовательности действий студентов. На втором этапе была поставлена и решена задача классификации: по агрегированным поведенческим признакам и метаинформации о студенте и курсе необходимо было предсказать, завершит ли студент курс.

Применение алгоритма PrefixSpan позволило получить частые поведенческие паттерны студентов. Однако, как показал анализ, различия между поведением студентов, завершивших курс, и тех, кто его бросил, **оказались менее выраженными, чем ожидалось**. Это свидетельствует о том, что одного лишь порядка действий в сессиях недостаточно для надёжного предсказания исхода обучения. Поведение пользователей внутри одной сессии может быть схожим, несмотря на разные финальные результаты. Тем не менее, выявленные последовательности дали ценную информацию о типичном поведении студентов и помогли при формировании признаков для модели классификации.

Во второй части работы была построена модель классификации с использованием ансамблевых алгоритмов — Random Forest и градиентного бустинга. Признаки включали количество каждого типа действия, длительность всех сессий, характеристики курса и студента. Полученные результаты показали высокую точность классификации: accuracy = 0.84, ROC-AUC = 0.74. Это подтверждает возможность прогнозирования отсева студентов на основе их поведения в системе.

После этого был проведён анализ важности признаков, а значит и действий пользователя, с использованием SHAP (SHapley Additive Explanations). Он позволил интерпретировать влияние отдельных признаков на итоговые предсказания модели.

Результаты SHAP визуализации позволили не только определить наиболее значимые признаки, но и понять, в какую сторону они влияют на предсказание (увеличивают или уменьшают вероятность отсева). Это усиливает доверие к модели и позволяет использовать её в практических целях.

Практическая значимость работы заключается в том, что построенная модель может быть интегрирована в платформы онлайн-обучения и использоваться для выявления студентов, находящихся в зоне риска отсева. Это откроет возможность для своевременных интервенций: отправки напоминаний, предоставления дополнительных ресурсов, подключения тьютора и т.д. Кроме того, выделенные паттерны поведения могут быть использованы для оптимизации структуры курсов.

## Список литературы

1. Han J., Pei J., Kamber M. Data Mining: Concepts and Techniques. – 3rd ed. – Morgan Kaufmann, 2011. – 744 с.
2. Müller A. C., Guido S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. – O’Reilly Media, 2016. – 400 с.
3. Kaggle. MOOC Dataset [Электронный ресурс]. – Режим доступа: <https://www.kaggle.com/datasets/samyakjhaveri/mooc-final/data>, свободный. – Дата обращения: 07.05.2025.
4. Apache Software Foundation. PySpark Documentation [Электронный ресурс]. – Режим доступа: <https://spark.apache.org/docs/latest/api/python/>, свободный. – Дата обращения: 07.05.2025.
5. Scikit-learn. Документация по библиотеке машинного обучения Scikit-learn [Электронный ресурс]. – Режим доступа: <https://scikit-learn.org/stable/index.html>, свободный. – Дата обращения: 01.05.2025.
6. SHAP. Документация по SHapley Additive Explanations [Электронный ресурс]. – Режим доступа: <https://shap.readthedocs.io/>, свободный. – Дата обращения: 09.05.2025.
7. Pandas. Документация по библиотеке pandas для анализа данных [Электронный ресурс]. – Режим доступа: <https://pandas.pydata.org/docs/>, свободный. – Дата обращения: 01.05.2025.

## Приложение

