Здравствуйте, уважаемые председатели и члены комиссии!

Меня зовут Ерёменко Артём Геннадьевич, группа НПИбд-01-18. Я хочу Вам представить свою ВКР на тему «Разработка информационной системы для анализа активности пользователей интернет-портала онлайн-образования».

Данная тематика актуальна из-за возросшей необходимости разработки средств онлайн-образования в связи со сложившейся в мире сложной эпидемиологической обстановкой, из-за которой достаточно внушительная часть студентов утратила возможность проходить очное обучение непосредственно в учебном заведении. Также актуальность связана с популярностью платформ для прохождения курсов по различным направлениям подготовки. Работа посвящена решению современной задачи по разработке Информационной Системы интернет-портала онлайн-образования, а также анализу активности пользователей с целью выявления закономерностей их поведения. Для анализа активности пользователей применяются анализ данных, методы разведочного анализа данных и машинного обучения с учителем.

На этом слайде вы можете ознакомиться со структурой ВКР. В первой главе ставятся задачи разработки Информационной системы и модуля анализа активности её пользователей, а также проводится обзор унифицированного языка моделирования с последующим построением диаграмм. Во второй главе проводятся изучение, описание и сравнение современных методов машинного обучения. В третьей главе разрабатываются методы решения задачи. В четвёртой главе проводятся эксперименты в рамках исследования.

Здесь вы можете увидеть пользователей Информационной Системы и действия, которые они могут выполнять. Посетитель – это неавторизованный пользователь сайта. Он может: Просмотреть информацию о курсах, Просмотреть информацию об организации, Обратиться в службу поддержки, Авторизоваться. Студент – это авторизованный пользователь системы с учебным аккаунтом. Он может: Купить курс, Открыть доступный курс, Просмотреть свой профиль, а также сделать всё то, что может и Посетитель. Администратор – это авторизованный пользователь системы с аккаунтом администратора. Он может: управлять курсами, Управлять пользователями, Прочитать обращения из службы поддержки. Последний тип пользователей системы – это преподаватель, который может проверять домашние задания студентов, просматривать свой профиль, а также, с предоставлением прав доступа, может выполнять обязанности администратора.

Диаграммы были построены при помощи UML. UML – это унифицированный графический язык моделирования для описания, визуализации, проектирования и документирования объектно-ориентированных систем. UML служит для поддержки процессов моделирования программных систем на основе объектно-ориентированного подхода, для организации взаимосвязи концептуальных и программных понятий, для отражения проблем масштабирования сложных систем. В ходе выполнения ВКР были разработаны диаграммы вариантов использования для каждого типа пользователя. Они повторяют информацию, изложенную мной ранее, поэтому не будем задерживаться на этом слайде.

Диаграммы классов демонстрируют иерархию классов и реализацию интерфейсов.

Диаграммы последовательности были созданы для трёх отдельных вариантов использования: Авторизоваться, Изменить курс и Проверить домашнее задание. Диаграммы Отражают динамическую составляющую системы и Порядок передачи сообщений.

Перейдём к практическому исследованию решения задачи анализа активностей пользователя. Данные для обработки были взяты из базы данных онлайн-университета Skillbox – это 5 таблиц со следующей информацией: таблица идентификаторов с названиями курсов; таблица с информацией о студентах; таблица с содержанием курсов; таблица для соотношения прогресса, студента и курса; таблица с фазами прохождения курсов.

В ходе обработки данных была построена диаграмма, демонстрирующая количество студентов на каждом направлении подготовки, а также рассчитаны различные метрики вроде статистики по возрастам студентов на каждом курсе. Красной линией отмечено медианное количество людей на курсе.

Далее был произведён расчёт потенциальной нагрузки на преподавателей, результаты которого можно увидеть на следующих двух графиках: прирост и прогресс на курсах по месяцам. Оказалось, что резкий прирост количества новых пользователей и прогресса на курсах, а соответственно и пик нагрузки, приходится на зимние праздники. Видимо, у студентов в этот период появляется больше времени, чтобы заниматься самообразованием.

В последнюю очередь, была выявлена сезонность прохождения курсов и рассчитано медианное время выполнения домашней работы по месяцам с января по декабрь для каждого курса. На графике видно, что по всем курсам наблюдается общая тенденция сокращения времени выполнения в феврале и июле с последующим возрастанием в марте и августе соответственно. Предположительно, все наблюдаемые процессы связаны с сезонами отпусков.

Также была разработана метрика успеваемости, основанная на следующем принципе: среднее время выполнения модуля конкретным студентом сравнивается с со средне-медианным временем выполнения такого же объёма заданий среди всех студентов. В результате, по каждому изучаемому курсу, если время меньше или равно порогу, то у студента нет отставания по программе обучения и ставится метка класса «0», в ином случае – «1». Так как курсов, а соответственно и меток, может быть несколько, то берётся медианное значение с округлением вниз.

Следующим этапом был проведён разведочный анализ данных, в ходе которого была составлена корреляционная матрица и отобраны атрибуты с наибольшей корреляцией с таргетным атрибутом «status» и наименьшей корреляцией между собой. Однако уже на этом этапе наблюдается слабая корреляция со статусом.

Далее были составлены диаграммы зависимости количества людей со значением определённого бинарного атрибута от возраста. Вдоль осей всех графиков соблюдается определённое Соотношение между противоположными значениями при каждом возрасте: 25 к 75 % женатых и неженатых, 60 к 40 % у кого есть отставание и нет, 54 к 46 % у кого есть дети и нет. Выбросы на графиках отсутствуют. При поиске зависимостей на графиках аномалий выявлено не было.

Были построены 4 модели машинного обучения на основе следующих методов: Логистическая регрессия, Дерево решений, Метод Опорных векторов и CatBoost. У всех моделей довольно низкие показатели.

TP — true positive, классификатор верно отнёс объект к рассматриваемому классу.

TN — true negative, классификатор верно утверждает, что объект не принадлежит к рассматриваемому классу.

FP — false positive, классификатор неверно отнёс объект к рассматриваемому классу.

FN — false negative, классификатор неверно утверждает, что объект не принадлежит к рассматриваемому классу.

Из проделанной работы можно сделать следующие выводы:

1. Были разработаны UML диаграммы Информационной Системы интернет-портала онлайн-образования. 4 диаграммы вариантов использования, диаграмма классов и 3 диаграммы последовательности для отдельно взятых вариантов использования.
2. Было проведено сравнение методов машинного обучения
3. Был разработан модуль для анализа активности пользователей
4. Было проведено сравнение построенных моделей
5. Эксперименты показали, что модуль машинного обучения ещё требует доработки

Перед вами основные источники информации, использовавшиеся при подготовке работы.

Спасибо Вам за внимание!

А теперь я буду рад выслушать ваши вопросы и ответить на них.