МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ»

Разработка информационной системы для анализа активности пользователей интернет-портала онлайн-образования

Руководитель ВКР:
к.ф.-м.н., доцент кафедры ИТ,
Хачумов М.В.

Выполнил:
Еременко Артем Геннадьевич

Группа:
НПИбд-01-18

Москва, 2022 г.

Специфика работы

Актуальность темы

Актуальность задачи заключается в 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.

Цель работы

Данная дипломная работа посвящена решению современной задачи по разработке ИС интернет-портала онлайн-образования и анализу активности её пользователей с целью выявления закономерностей их поведения.

Методы исследования

В данной дипломной работе для решения задачи анализа активностей пользователей интернет-портала применяются анализ данных, методы EDA и машинного обучения с учителем.

Структура ВКР

I глава:

- постановка задачи разработки ИС
- постановка задачи анализа активности пользователей ИС
- обзор языка UML
- разработка основных диаграмм

II глава:

- исследование современных методов машинного обучения
- формальное описание изученных методов

III глава:

- предложение методов и алгоритмов решения поставленной задачи

IV глава:

- исследование практических вопросов решения задачи
- проведение экспериментальных исследований

Пользователи ИС и их возможности

Посетитель

1)Просмотреть информацию о курсах
2)Просмотреть информацию об организации
3)Обратиться в службу поддержки
4)Авторизоваться
 a)Войти в аккаунт
 b)Создать аккаунт

Студент

1)Купить курс
 a)Выбрать метод оплаты
  i)Выбрать оплату картой онлайн
  ii)Выбрать оплату платёжным сервисом
  iii)Оформить рассрочку на покупку
 b)Применить промо-код на скидку
2)Открыть доступный курс
 a)Перейти в модуль
  i)Открыть урок
  (1)Просмотреть обучающие видео урока
  (2)Сдать домашнюю работу
  (3)Написать курирующему преподавателю
3)Просмотреть свой профиль
 a)Редактировать личные данные
 b)Изменить пароль

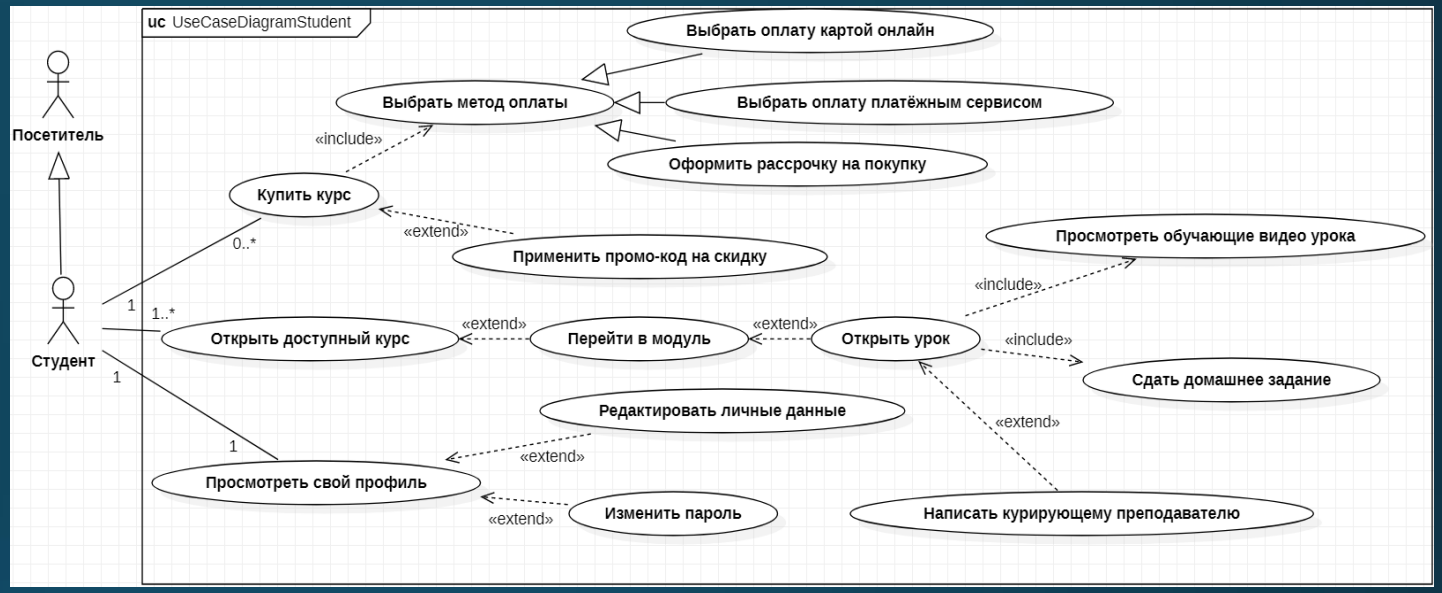
Преподаватель

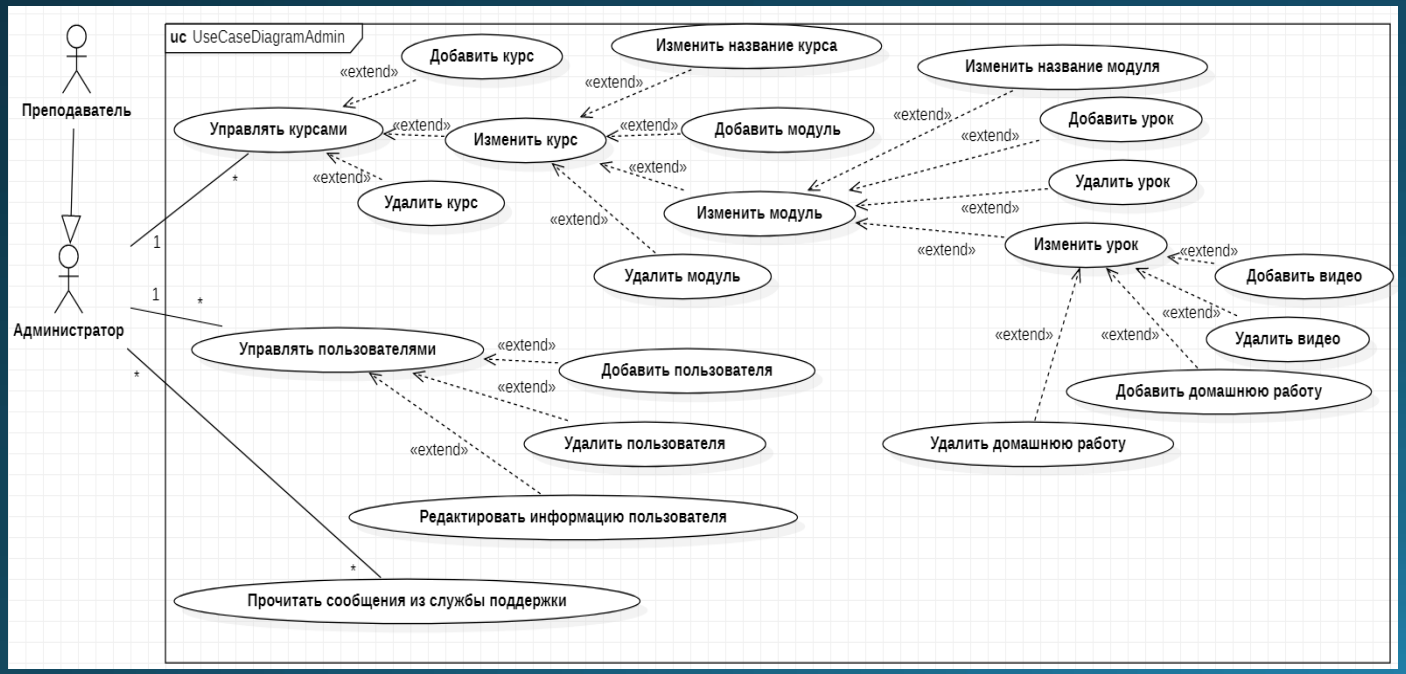
1)Проверить домашнее задание
 a)Вернуть домашнее задание
 b)Принять домашнее задание
 c)Оставить комментарий
2)Просмотреть свой профиль
 a)Редактировать личные данные
 b)Изменить пароль

Администратор

1)Управлять курсами
 a)Добавить курс
 b)Удалить курс
 c)Изменить курс
  i)Изменить название курса
  ii)Добавить модуль
  iii)Удалить модуль
  iv)Изменить модуль
   (1)Изменить название модуля
   (2)Добавить урок
   (3)Удалить урок
   (4)Изменить урок
    (a)Добавить видео
    (b)Удалить видео
    (c)Добавить домашнюю работу
    (d)Удалить домашнюю работу
2)Управлять пользователями
 a)Добавить пользователя
 b)Удалить пользователя
 c)Редактировать инф. пользователя
3)Прочитать сообщения службы поддержки

Диаграммы вариантов использования





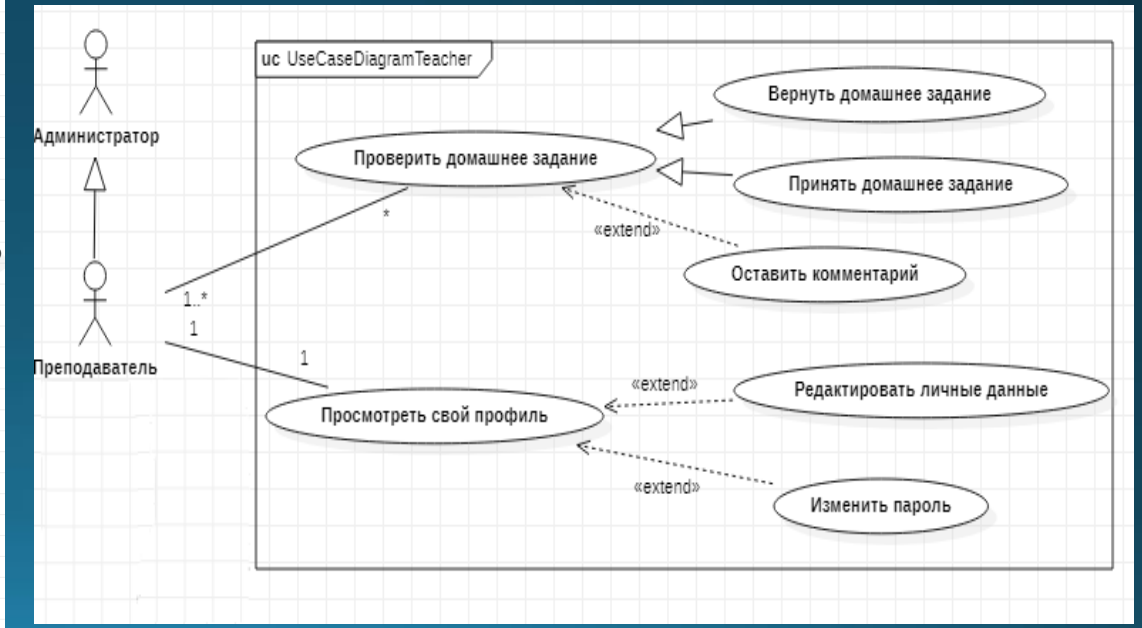
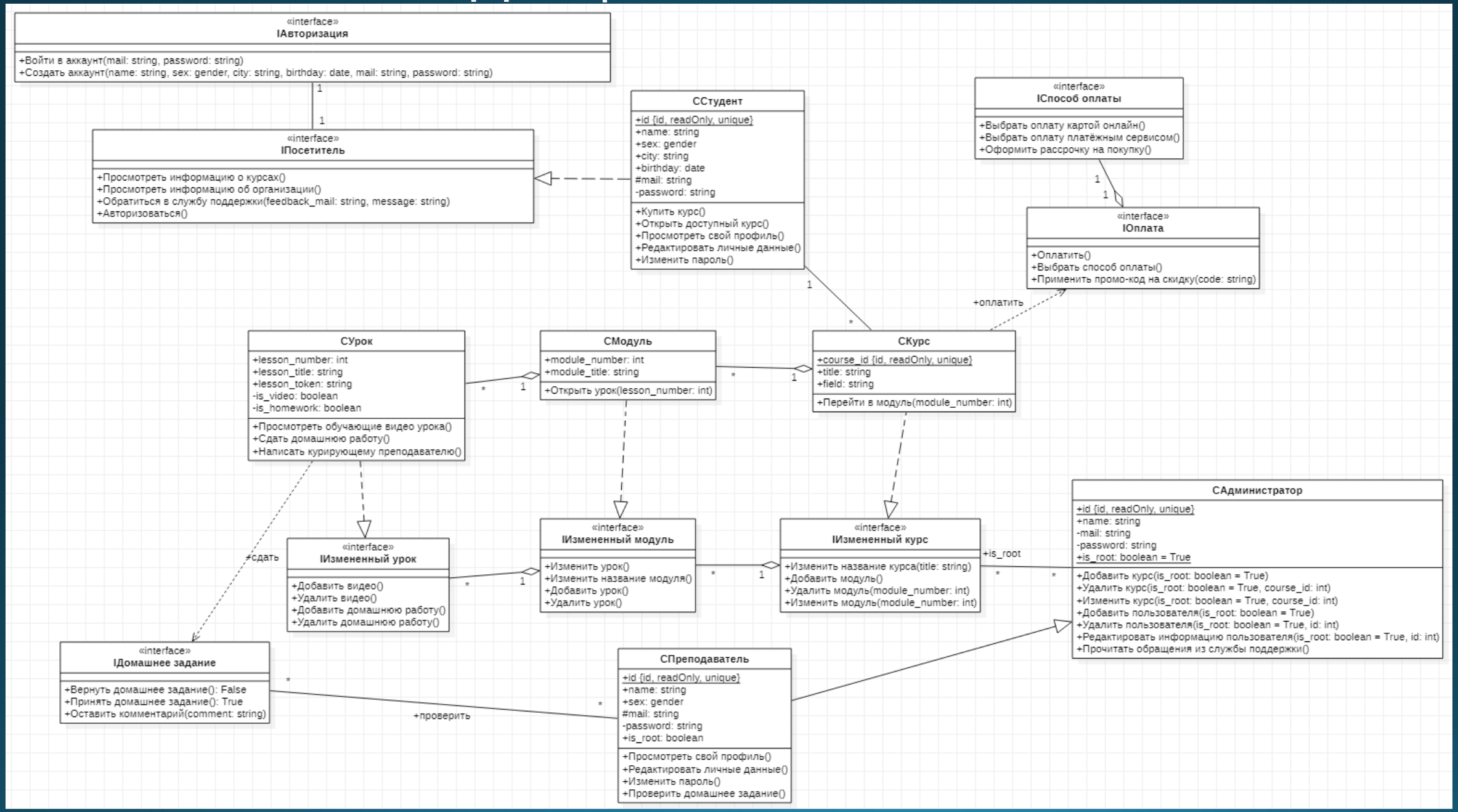
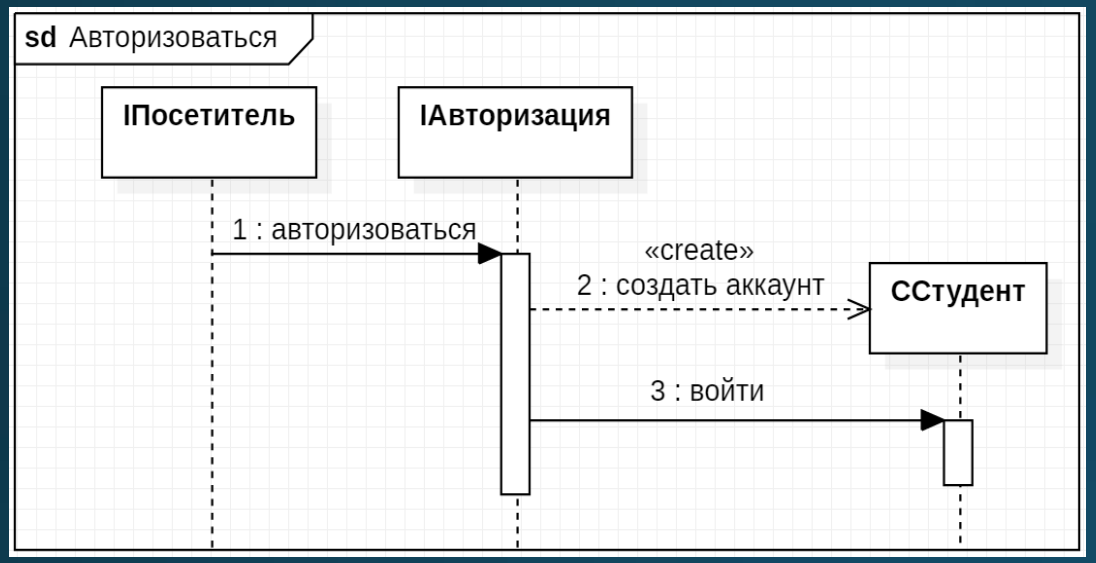
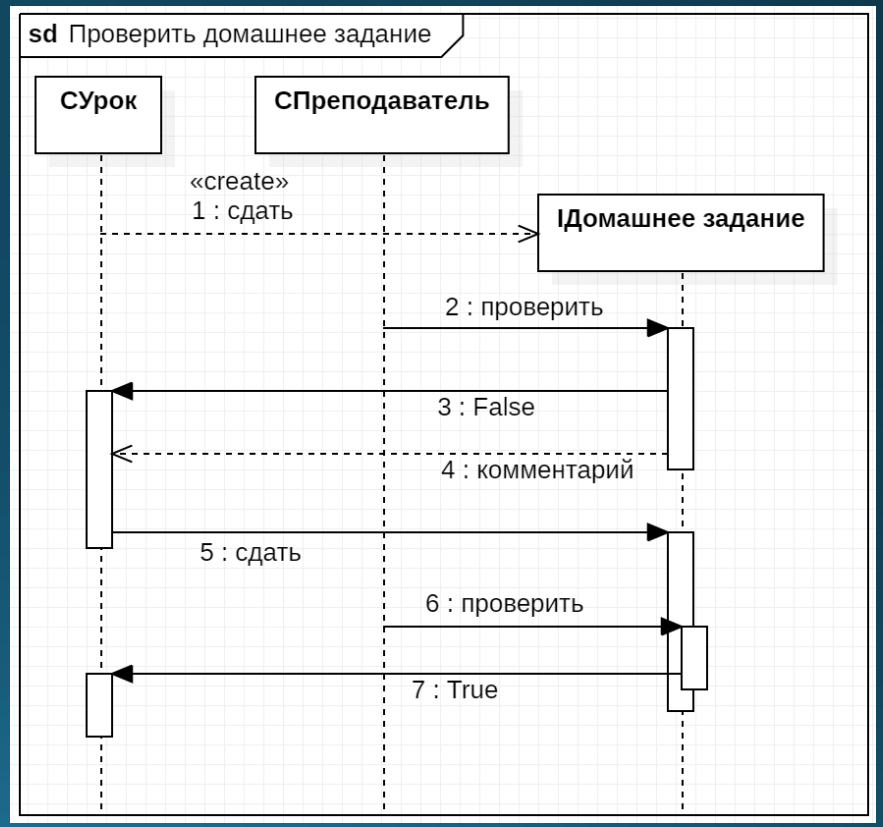


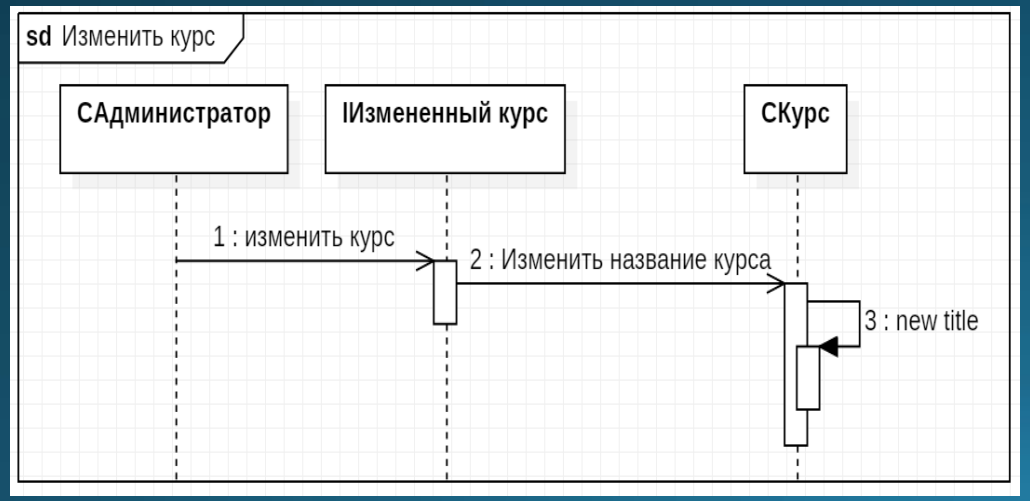
Диаграмма классов



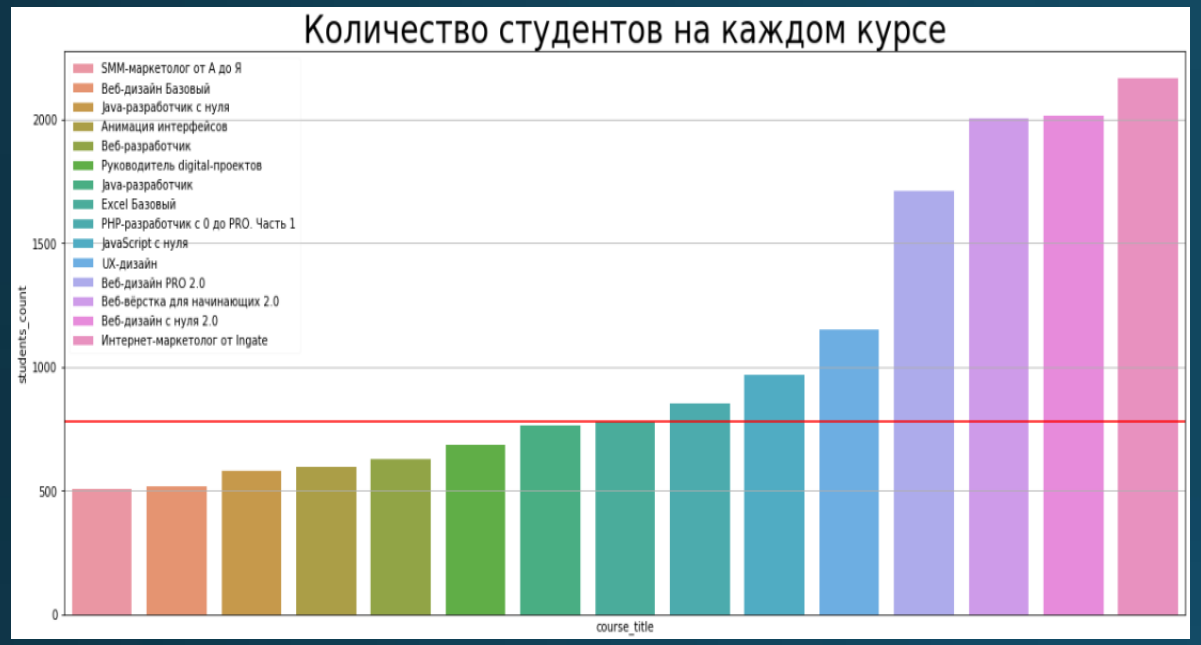
Диаграммы последовательности действий

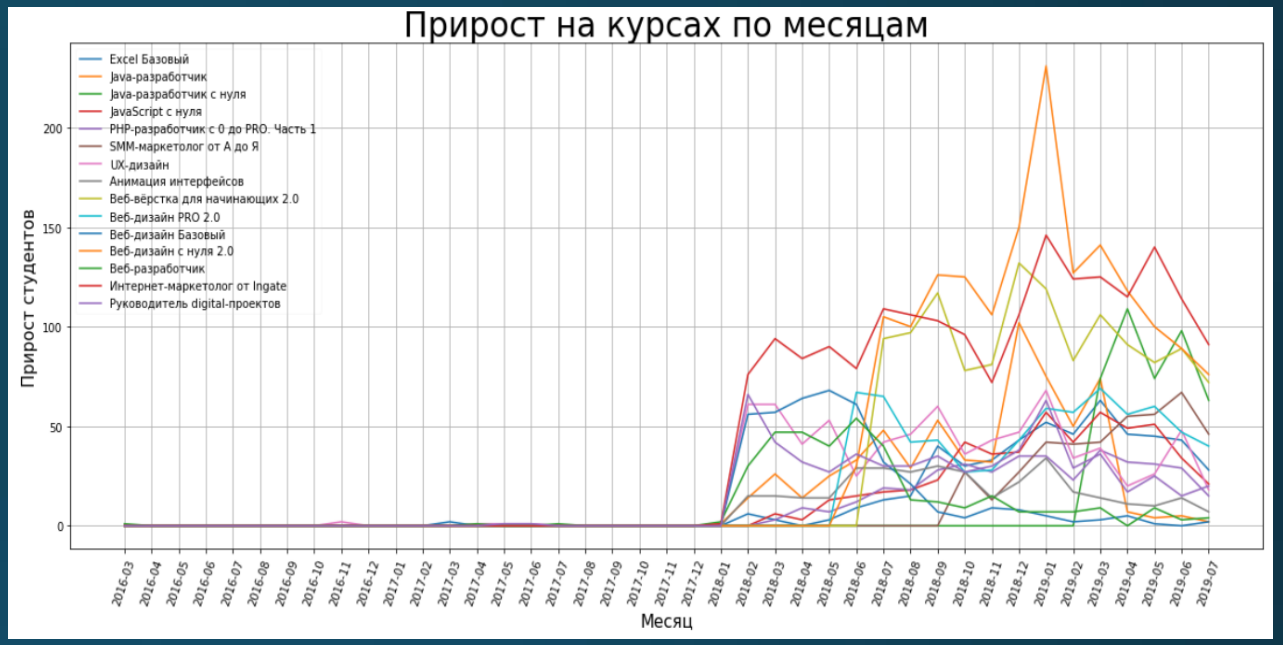


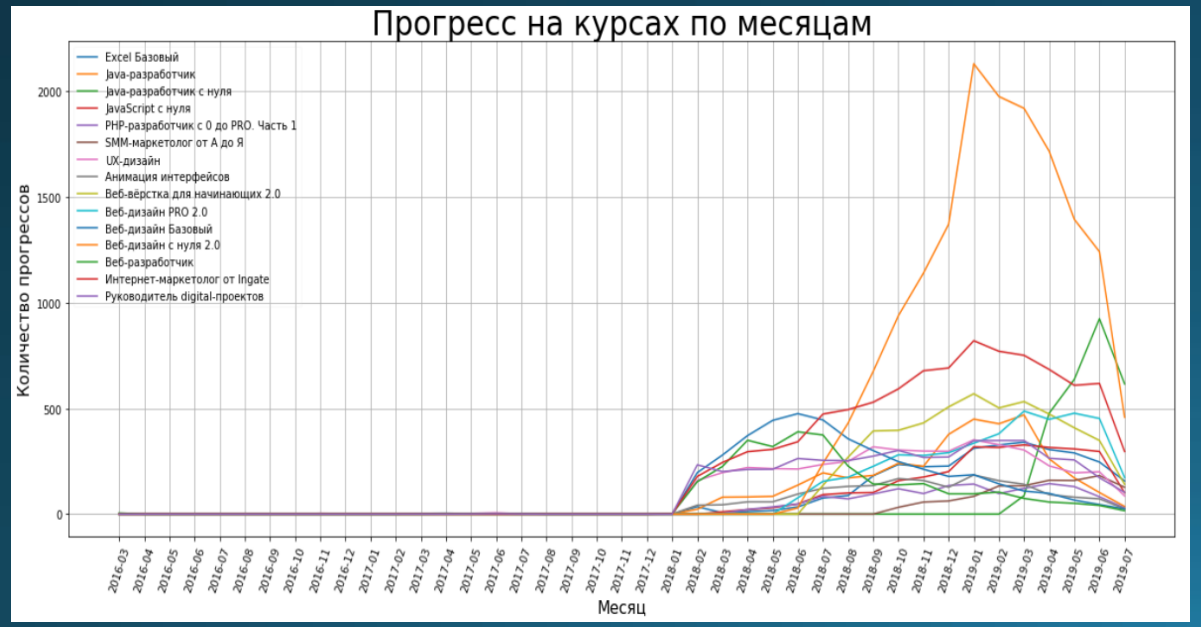




Практическое исследование решения задачи анализа активностей пользователя







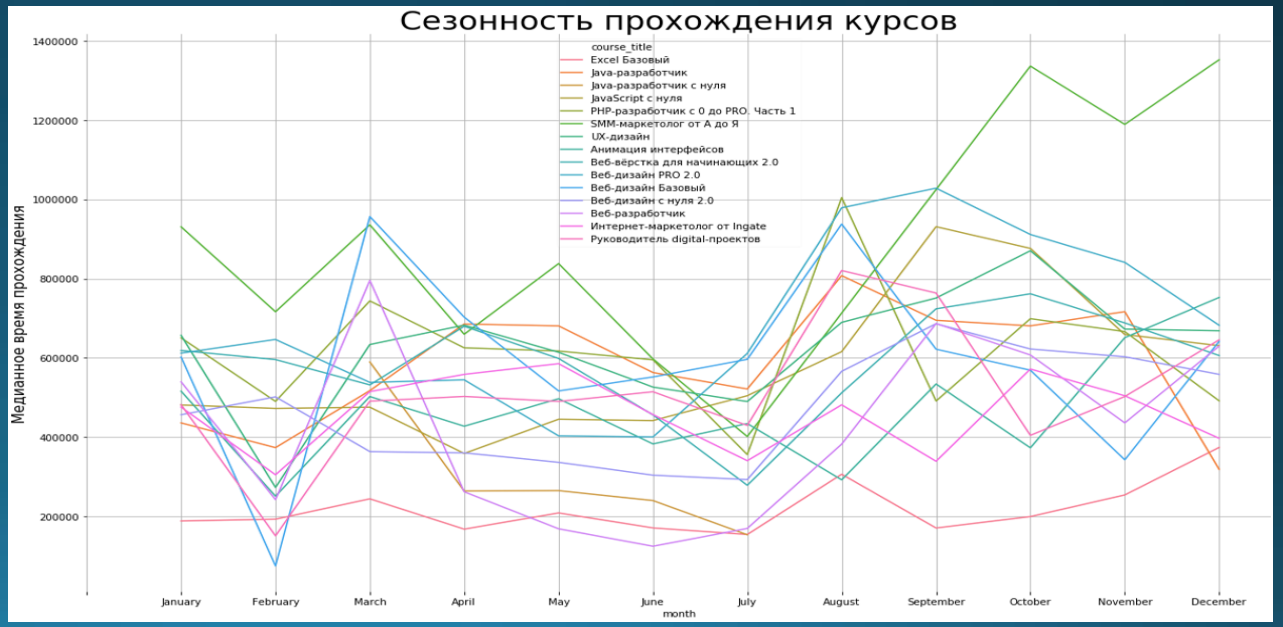


Таблица с метками успеваемости

```
# записываем вычисленные статусы
status_df['statuses'] = total_list
# приводим к медианным значениям с округлением вниз
status_df['status'] = [math.floor(statistics.median(x)) for x in status_df['statuses']]
status_df
```

	id	courses	statuses	status
0	768c2987a744c51ce64a5993a2a94eaf	[JavaScript с нуля, Анимация интерфейсов, Веб-...	[0, 0, 0]	0
1	03151bc73bdb29fe1be1443c6d83e22f	[UX-дизайн, Анимация интерфейсов, Веб-дизайн P...	[1, 1, 1, 1]	1
2	ed235f47e16da6e83d3f1cb511f38ea6	[Веб-дизайн PRO 2.0, Веб-дизайн с нуля 2.0]	[0, 1]	0
3	59e8681cb7b5c8043ae1aac10c8053ca	[Excel Базовый, Анимация интерфейсов, Веб-диза...	[0, 1, 1]	1
4	c16250079190337fe9074736e33eecb2	[Веб-дизайн PRO 2.0, Веб-дизайн с нуля 2.0]	[0, 1]	0
...	...	...	...	...
9230	a88d8e65143914ccc002c8abbe91324e	[Java-разработчик с нуля]	[1]	1
9231	5b9acd377d0d1b1f2e9e324a44dd0c8a	[Java-разработчик с нуля]	[0]	0
9232	71b5e788516d8e83fb9dc3b5f869dd5b	[Java-разработчик с нуля]	[0]	0
9233	0b77dc9de3ebc312a2ff105bef4b443b	[Java-разработчик с нуля]	[1]	1
9234	b90e440def5b7a643395eed52c02a339	[Java-разработчик с нуля]	[0]	0

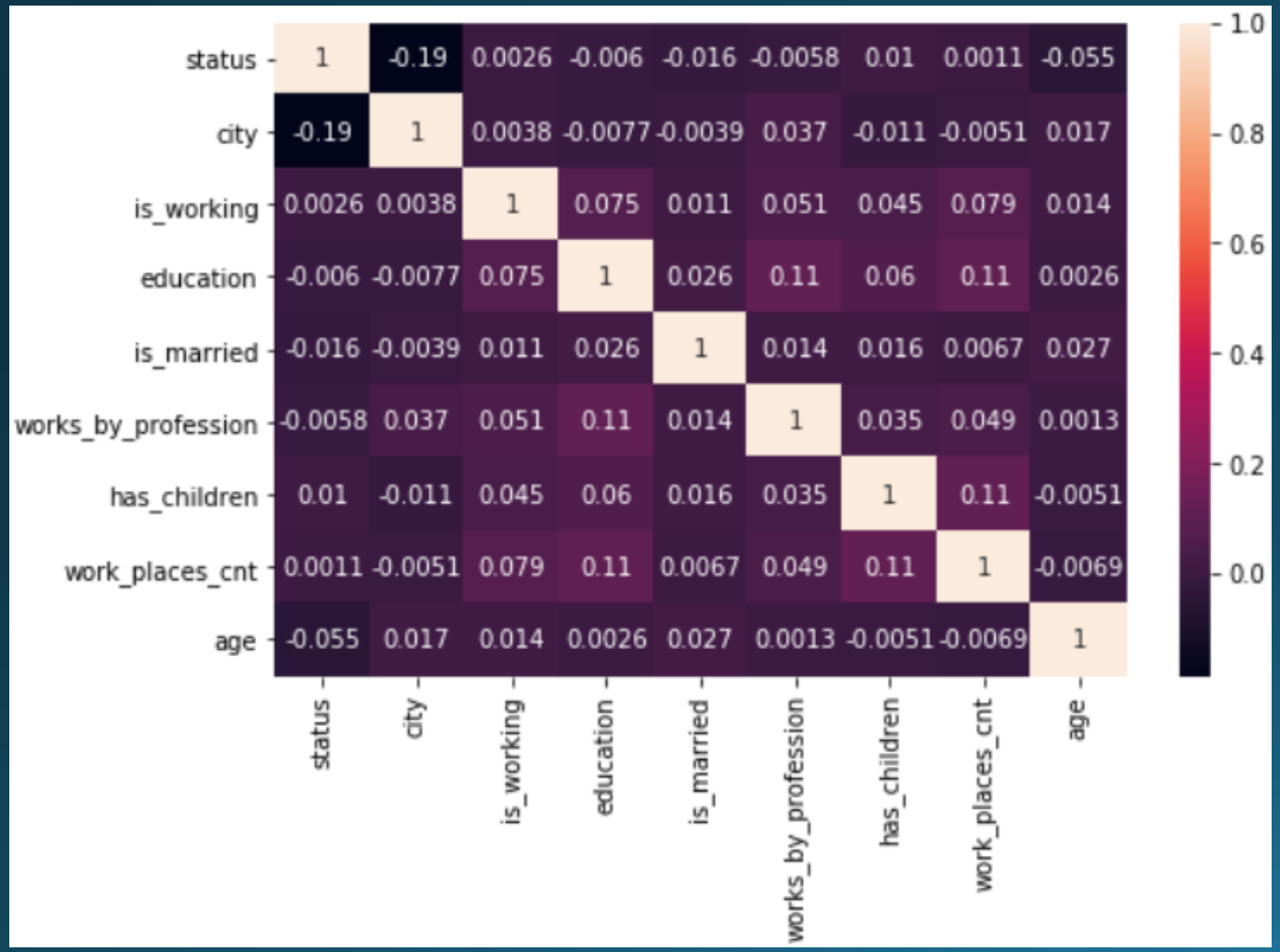
9235 rows × 4 columns

```
# соединяем таблицы
final_df = pd.merge(status_df,\
                    students,\
                    on='id')
# удаляем ненужные и промежуточные атрибуты
final_df.drop(columns=['statuses','courses','id_','birthday'], inplace=True)
final_df
```

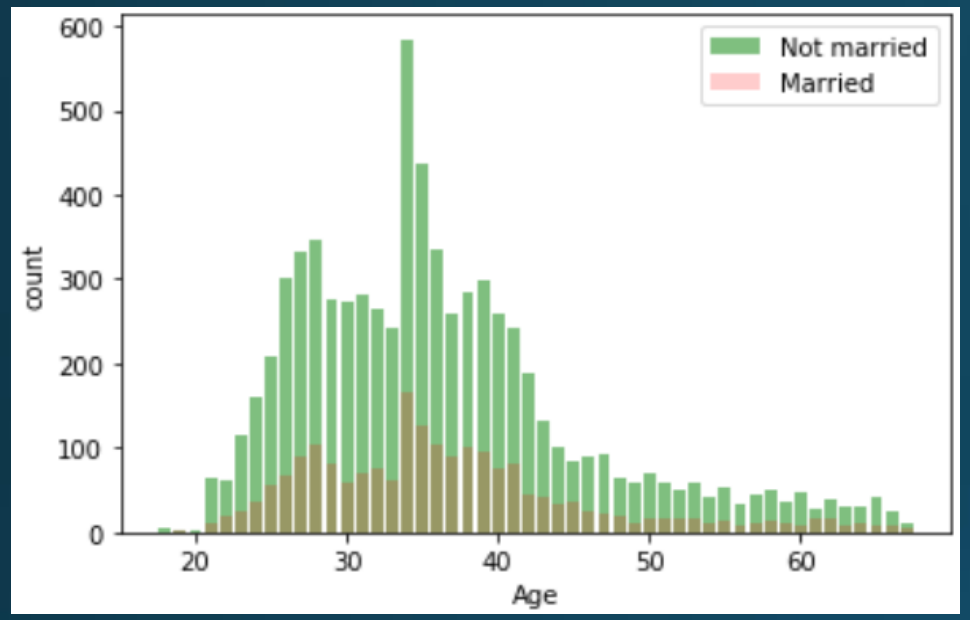
	id	status	city	is_working	education	is_married	works_by_profession	has_children	work_places_cnt	age
0	768c2987a744c51ce64a5993a2a94eaf	0	Санкт-Петербург	True	3	False	True	True	5	31
1	03151bc73bdb29fe1be1443c6d83e22f	1	Санкт-Петербург	True	3	True	False	False	6	27
2	ed235f47e16da6e83d3f1cb511f38ea6	0	Москва	True	2	True	False	False	6	30
3	59e8681cb7b5c8043ae1aac10c8053ca	1	Самара	True	3	False	True	False	5	26
4	c16250079190337fe9074736e33eecb2	0	Москва	True	2	False	False	True	5	27
...	...	...	...	...	...	...	...	...	...	...
9230	a88d8e65143914ccc002c8abbe91324e	1	Томск	True	2	True	False	True	4	37
9231	5b9acd377d0d1b1f2e9e324a44dd0c8a	0	Lausanne	True	2	False	False	True	6	47
9232	71b5e788516d8e83fb9dc3b5f869dd5b	0	Омск	False	2	True	True	False	6	35
9233	0b77dc9de3ebc312a2ff105bef4b443b	1	Тимашевск	False	3	False	True	False	5	52
9234	b90e440def5b7a643395eed52c02a339	0	Н.Новгород	True	2	True	True	True	2	45

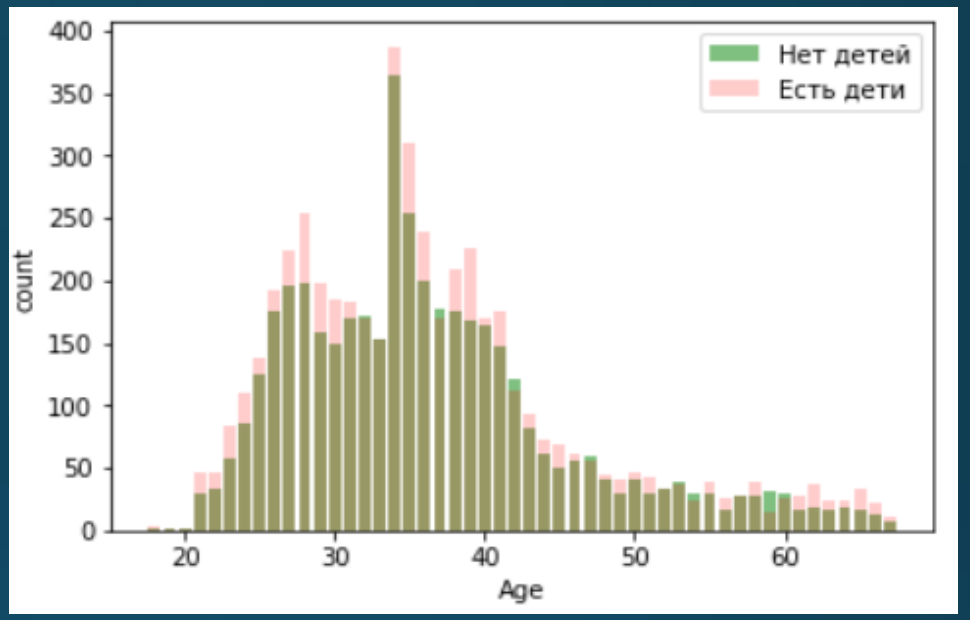
9235 rows × 10 columns

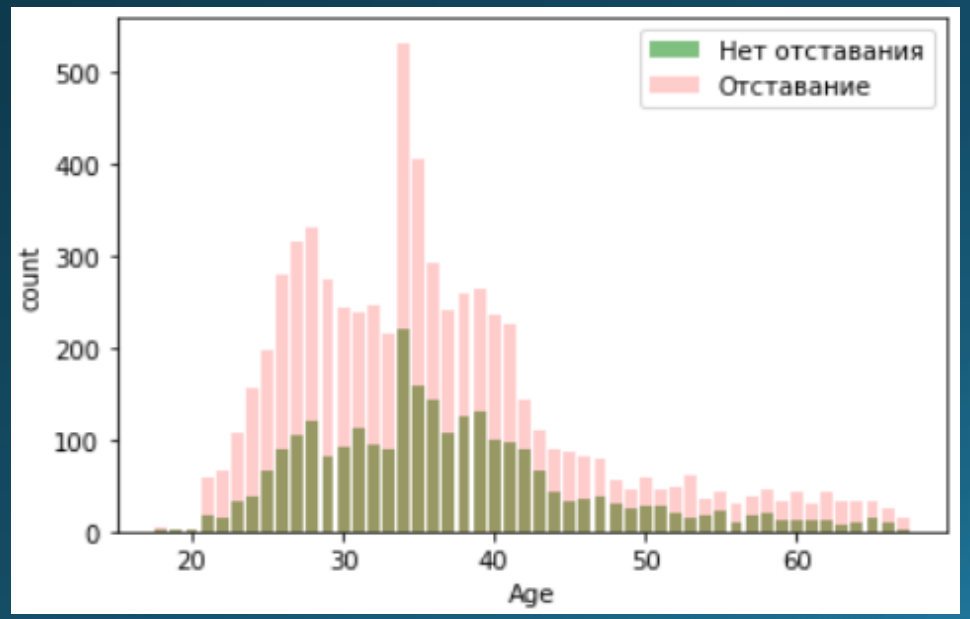
Разведочный анализ данных

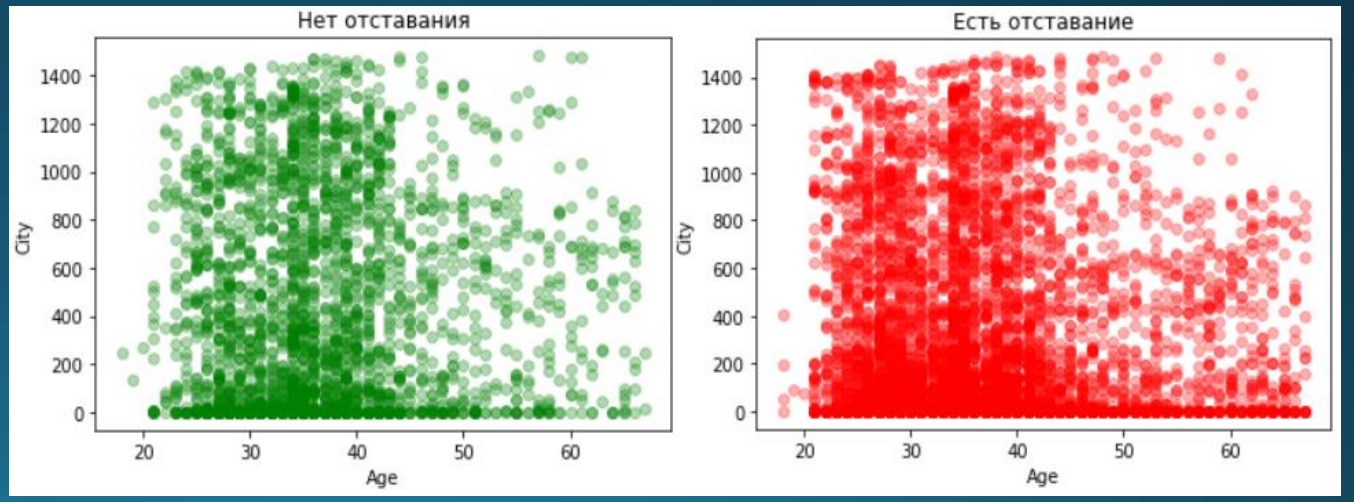


Разведочный анализ данных ч.2









Результаты обучения моделей

Таблица 4.3 – Оценки точности методов машинного обучения

Метод	TP	TN	FP	FN	Accuracy	Precision	Error rate
Логистическая регрессия	19	1616	646	28	0,708098744	0,02857143	0,2919013
Дерево решений	166	1317	499	327	0,642269381	0,24962406	0,3577306
SVM	0	1644	655	0	0,715093519	0	0,2849065
CatBoost	45	1583	620	61	0,705067129	0,06766917	0,2949329

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FN}$$

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN}$$

Выводы

- Разработаны UML диаграммы ИС интернет-портала онлайн-образования
- Проведено сравнение методов машинного обучения
- Разработан модуль для анализа активности пользователей
- Проведено сравнение построенных моделей машинного обучения
- Эксперименты показали, что модуль машинного обучения требует доработки

Основные источники

- Якобсон А., Буч Г., Рамбо Дж. Краткая история UML // Язык UML. Руководство пользователя = The Unified Modeling Language Usere Guide. 2-е. – М.: ДМК Пресс, 2006. – 496 с.
- Кватрани Т. Rational Rose 2000 и UML. Визуальное моделирование: Пер. с англ. – М.: ДМК Пресс, 2013. – 176 с.
- Рашка С. Python и машинное обучение. / пер. с англ. А.В. Логунова. – М.: ДМК Пресс, 2017. – 418 с.
- Дейтел Пол, Дейтел Харви. Python: Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.: ил. — (Серия «Для профессионалов»)
- Rudolph Russell. Machine Learning: Step-by-Step Guide to Implement Machine Learning Algorithms with Python. – CreateSpace Independent Publishing Platform, 2018. – 106 с.
- Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. — Cambridge University Press, 2000. – 204 с.
- Juliana Tolles, William J. Meurer. Logistic Regression: Relating Patient Characteristics to Outcomes. // JAMA, 2016. – Т. 316, № 5. – С. 533-534.

Приложение

```
# создаём датафрейм, в котором собрали время выполнения всех заданий
df_metric = table[(table['status']=='done')][['course_title',\
                                              'module_number',\
                                                 'lesson_number',\
                                                 'student_id',\
                                              'start_date',\
                                              'finish_date']]\
                                                          .reset_index(drop=True)
```

```
# преобразуем строковый тип даты к типу  datetime.datetime
df_metric['start_date'] = [datetime.datetime.strptime(df_metric['start_date'][i][0:18],\
                                                      "%Y-%m-%d %H:%M:%S") for i in df_metric.index]

# преобразуем строковый тип даты к типу  datetime.datetime
df_metric['finish_date'] = [datetime.datetime.strptime(df_metric['finish_date'][i][0:18],\
                                                       "%Y-%m-%d %H:%M:%S") for i in df_metric.index]

# вычисляем разность между временем начала и окончания прохождения урока
df_metric['spent_time'] = [relativedelta(df_metric['finish_date'][i],\
                                         df_metric['start_date'][i]) for i in df_metric.index]
# переводим затраченное время в секунды
df_metric['spent_sec'] = [(df_metric['spent_time'][x].days*24*60*60\
                           +df_metric['spent_time'][x].hours*60*60\
                           +df_metric['spent_time'][x].minutes*60\
                           +df_metric['spent_time'][x].seconds) for x in range(len(df_metric))]
```

```
df_metric.head()
```

	course_title	module_number	lesson_number	student_id	start_date	finish_date	spent_time	spent_sec
0	Веб-дизайн PRO 2.0	2	4	768c2987a744c51ce64a5993a2a94eaf	2018-06-23 08:28:05	2018-06-23 08:28:05	relativedelta()	0
1	Веб-дизайн PRO 2.0	2	4	03151bc73bdb29fe1be1443c6d83e22f	2019-03-18 14:23:01	2019-03-18 14:54:05	relativedelta(minutes=+31, seconds=+4)	1864
2	Веб-дизайн PRO 2.0	2	4	ed235f47e16da6e83d3f1cb511f38ea6	2019-07-09 09:18:04	2019-07-11 08:03:00	relativedelta(days=+1, hours=+22, minutes=+44,...	168296
3	Веб-дизайн PRO 2.0	2	4	59e8681cb7b5c8043ae1aac10c8053ca	2018-07-27 15:39:01	2018-07-27 16:13:05	relativedelta(minutes=+34, seconds=+4)	2044
4	Веб-дизайн PRO 2.0	2	4	c16250079190337fe9074736e33eecb2	2019-04-24 18:42:04	2019-04-24 18:44:05	relativedelta(minutes=+2, seconds=+1)	121

```
# вычисляем среднее время, которое конкретный студент тратит на выполнение модуля
df_metric_task_1 = pd.DataFrame(df_metric.groupby(['course_title', 'module_number','student_id'])['spent_sec'].mean())
df_metric_task_1.rename(columns={'spent_sec':'mean'}, inplace=True)
df_metric_task_1.reset_index(['course_title', 'module_number','student_id'], inplace=True)
```

```
# вычисляем медианное время, которое студенты тратят на выполнение каждого урока
df_metric_task_2 = pd.DataFrame(df_metric.groupby(['course_title', 'module_number', 'lesson_number'])['spent_sec'].median())
df_metric_task_2.rename(columns={'spent_sec':'median'}, inplace=True)
df_metric_task_2.reset_index(['course_title', 'module_number', 'lesson_number'], inplace=True)
```

```
# вычисляем среднее время, которое студенты тратят на выполнение каждого модуля
df_metric_task_3 = pd.DataFrame(df_metric_task_2.groupby(['course_title', 'module_number'])['median'].mean())
df_metric_task_3.rename(columns={'median':'mean_median'}, inplace=True)
df_metric_task_3.reset_index(['course_title', 'module_number'], inplace=True)
```

```
df_metric_task_3
```

	course_title	module_number	mean_median
0	Excel Базовый	1	24517.000000
1	Excel Базовый	2	23743.785714
2	Excel Базовый	3	54148.833333
3	Excel Базовый	4	43727.071429
4	Excel Базовый	5	51496.000000
...	...	...	...
224	Руководитель digital-проектов	13	207870.250000
225	Руководитель digital-проектов	14	90706.357143
226	Руководитель digital-проектов	15	18186.958333
227	Руководитель digital-проектов	16	73893.875000
228	Руководитель digital-проектов	17	4200.250000

229 rows × 3 columns

```python
# создаём список студентов, которые приобрели хотя бы один курс
stud_paid_list = list(df_metric['student_id'].unique())
# формируем датафрейм для статусов студентов
status_df = pd.DataFrame(stud_paid_list, columns=['id'])
# записываем курсы, которые у них есть
status_df['courses'] = [[y for y in df_metric_task_1[df_metric_task_1['student_id']==x]['course_title'].unique()] for x in stud_p
```

```python
total_list = []
list_len = len(stud_paid_list)
# проходим посписку студентов с оплаченными курсами
for i in stud_paid_list:
    # проходим по всем курсам, что есть у этого студента
    for j in status_df[status_df['id']==i]['courses']:
        curr_list = []
        # находим все модули, которые студент уже выполнил
        for k in j:
            modules_num = list(df_metric_task_1[(df_metric_task_1['student_id']==i)\
                                                &(df_metric_task_1['course_title']==k)]['module_number'].values)
            summary = 0
            # суммируем средние медианные значения времени по пройденным модулям
            for l in modules_num:
                summary += df_metric_task_3[(df_metric_task_3['course_title']==k)\
                                            &(df_metric_task_3['module_number']==l)]['mean_median'].values

            # вычисляем среднее средних значений времени выполнения модулей
            curr_mean = df_metric_task_1[(df_metric_task_1['student_id']==i)\
                                         &(df_metric_task_1['course_title']==k)]['mean'].mean()
            # вычисленные значения между собой
            if summary/len(modules_num) >= curr_mean:
                curr_list.append(0)
            else:
                curr_list.append(1)

        total_list.append(curr_list)
```

```python
# записываем вычисленные статусы
status_df['statuses'] = total_list
# приводим к медианным значениям с округлением вниз
status_df['status'] = [math.floor(statistics.median(x)) for x in status_df['statuses']]
status_df
```

	id	courses	statuses	status
0	768c2987a744c51ce64a5993a2a94eaf	[JavaScript с нуля, Анимация интерфейсов, Веб-...	[0, 0, 0]	0
1	03151bc73bdb29fe1be1443c6d83e22f	[UX-дизайн, Анимация интерфейсов, Веб-дизайн P...	[1, 1, 1, 1]	1
2	ed235f47e16da6e83d3f1cb511f38ea6	[Веб-дизайн PRO 2.0, Веб-дизайн с нуля 2.0]	[0, 1]	0
3	59e8681cb7b5c8043ae1aac10c8053ca	[Excel Базовый, Анимация интерфейсов, Веб-диза...	[0, 1, 1]	1
4	c16250079190337fe9074736e33eecb2	[Веб-дизайн PRO 2.0, Веб-дизайн с нуля 2.0]	[0, 1]	0
...	...	...	...	...
9230	a88d8e65143914ccc002c8abbe91324e	[Java-разработчик с нуля]	[1]	1
9231	5b9acd377d0d1b1f2e9e324a44dd0c8a	[Java-разработчик с нуля]	[0]	0
9232	71b5e788516d8e83fb9dc3b5f869dd5b	[Java-разработчик с нуля]	[0]	0
9233	0b77dc9de3ebc312a2ff105bef4b443b	[Java-разработчик с нуля]	[1]	1
9234	b90e440def5b7a643395eed52c02a339	[Java-разработчик с нуля]	[0]	0

9235 rows × 4 columns