ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА СЫРЫХ ДАННЫХ О ХАРАКТЕРИСТИКАХ СТУДЕНТОВ УНИВЕРСИТЕТА И РАЗВЕДОЧНЫЙ АНАЛИЗ

Е.И. Губин., к.ф.-м.н., доцент ОИТ ИШИТР
А. В. Семенюта
Томский политехнический университет
E-mail: avs183@tpu.ru

Введение

Одним из важнейших направлений стратегии развития любого университета является повышение качества образования. В числе прочих подходов используется интеллектуальный анализ накопленных данных и сведений.

Целью данной работы является предобработка дынных о студентах Томского Политехнического Университета на весеннюю сессию 2019 года и проведение разведочного анализа.

Для предобработки и анализа данных использовался скриптовый язык программирования Python и среды разработки Google Collab и Jupyter Lab. Результаты данной работы использованы для построения предсказательных моделей.

Описание набора данных

Используемый набор данных состоит из 8551 записи студентов и 26 столбцов-характеристик обучающихся (рис. 1)

```
RangeIndex: 8551 entries, 0 to 8550
Data columns (total 26 columns):
 #   Column                                                  Non-Null Count  Dtype  
---  ------                                                  --------------  -----  
 0   Форма обучения                                          8551 non-null   object 
 1   Квалификация                                            8551 non-null   object 
 2   Курс                                                    8551 non-null   int64  
 3   Специальность                                           8551 non-null   object 
 4   Профиль                                                 6676 non-null   object 
 5   Выпуск. отдел.                                          8551 non-null   object 
 6   Выпуск. школа                                           7988 non-null   object 
 7   Группа                                                  8551 non-null   object 
 8   Обуч. подразд.                                          8551 non-null   object 
 9   Фамилия                                                 8551 non-null   object 
 10  Имя                                                     8551 non-null   object 
 11  Отчество                                                8551 non-null   object 
 12  Форма финансирования                                    8551 non-null   object 
 13  Страна                                                  8533 non-null   object 
 14  Гражданство                                             8551 non-null   object 
 15  Пол                                                     8551 non-null   object 
 16  Дата рождения                                           8551 non-null   object 
 17  Академ отпуск (действующий) - да / нет                  8551 non-null   object 
 18  Всего                                                   8551 non-null   int64  
 19  Положительных                                           8551 non-null   int64  
 20  Неудовлетворительных                                    8551 non-null   int64  
 21  Дисциплины по которым получены неудовлетворительные оценки  6370 non-null   object 
 22  Пропусков по дисциплинам по которым получены неудовлетворительные оценки  8551 non-null   int64  
 23  Всего часов по дисциплинам по которым получены неудовлетворительные оценки  8539 non-null   float64
 24  Всего часов пропусков в семестре                        8551 non-null   int64  
 25  Всего часов аудиторных занятий в семестре               8470 non-null   float64
dtypes: float64(2), int64(6), object(18)
memory usage: 1.7+ MB
```

Рис. 1. Исследуемый набор данных

Предобработка данных.

Предобработка данных была проведена следующим образом. Столбцы «Фамилия», «Имя», «Отчество» были объединены в новый столбец «Полное имя», каждому студенту присвоен индекс, получившиеся данные вынесены в отдельный набор.

Также были удалены все факультативы из рассмотрения, так как они портили статистику другим студентам.

Так как в наборе данных нет переменной, значения которой характеризует успешность студента, было решено создать числовую переменную «Успешность», построенную путем деления столбца «Всего» на столбец «Положительных».

Далее все студенты были разбиты на три класса. В класс «0» записаны студенты, чья успешность ≤ 0.25. В класс «1» – от 0.25 до 0.75, не включая граничные значения. Чья успешность ≥ 0.75 – записаны в класс «2».

Далее была рассмотрена таблица корреляции столбцов данного набора признаков (рисунок 3). Будем считать, что признаки линейно зависимы, если значение коэффициента корреляции ≥ 0.75 по модулю.

Заметим, что имеются линейные зависимости между целевыми признаками «Класс» и «Успешность», что подчеркивает построение одного признака на основе другого. Следующая зависимость – между признаками «Форма Обучения_Очная» и «Всего аудиторных занятий в семестре». Других линейных зависимостей нет.

После удаления столбцов, не представляющих интереса для дальнейшего анализа, осталось 15 признаков. Все признаки были нормализованы.

Разведочный анализ итогового набора данных

Проведем разведочный анализ итогового набора данных. На рисунке 2 представлена гистограмма распределения студентов по несданным дисциплинам. Заметно, что студентов, не имеющих долгов – большинство.

Рис. 2. Словарь частот несданных дисциплин

Самая часто встречаемая несданная дисциплина – «Учебно-исследовательская работа студентов». Возможно, студенты недооценивают этот предмет. Также можно допустить, что студенты несерьезно относятся к предметам «Прикладная физическая культура» и «Иностранный язык (английский)», считая, что следует уделить внимание более сложным дисциплинам.

В ходе дальнейшего исследования выяснилось, что среди студентов большинство родилось в промежуток с 1997-го по 2000-ый год. Большинство студентов обучается очно на бакалавриате. Форма финансирования – бюджетная. Большинство студентов являются мужчинами и обучаются в Инженерной школе природных ресурсов и Инженерной школе энергетики. Большинство обучающихся в указанный период в академическом отпуске не находились.

Среднее арифметическое количество несданных дисциплин находится в районе двух, а медиана равна 3.28. Среднеквадратичное отклонение равно 3.36 несданных дисциплин.

Сформулируем статистические гипотезы. На рисунке 3 представлена таблица сопряженности класса студента и его пола. Заметно, что в каждом из классов мужчин больше, чем женщин, однако во втором классе процентное отношение женщин резко возрастает. Возможно, это повлияет на прогнозирование успешности студента.

Класс	0	1	2	All
Пол				
Женский	477	416	1608	2501
Мужской	1637	1429	2984	6050
All	2114	1845	4592	8551

Рис. 3. Таблица сопряженности класса студента и его пола

Проверим, являются ли распределения успешности мужчин и женщин нормальными. На рисунке 4 приведены соответствующие Q-Q графики. По графика видно, что распределения далеки от нормальных.

Для большей точности проверим нормальность критерием Шапиро-Уилка. При взятом уровне значимости α=0.05 p-значения и в том, и другом случаях равны нулю, следовательно, отвергаем нулевые гипотезы, что распределения нормальны. Такая ситуация повторяется для любого признака. Вид распределения не определен, поэтому придется использовать непараметрические критерии. Мы сравниваем две независимые выборки, поэтому будет использовать критерий Манна-Уитни и перестановочный критерий. Сформулируем первую гипотезу: «Средняя успешность студентов-женщин значимо выше, чем успешность студентов-мужчин». Уровень значимости примем за 0.05.

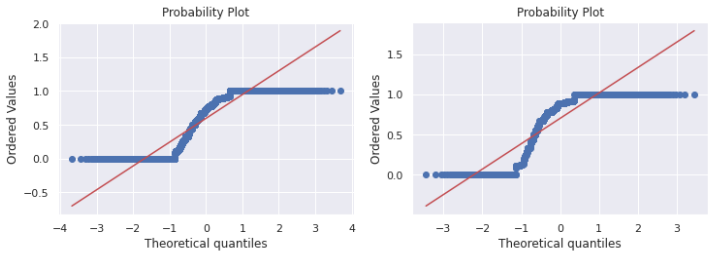

Рис. 4. Q-Q графики распределения успешности мужчин и женщин

Применяя аппарат формулирования и проверки статистических гипотез, получаем следующие выводы:

1. Студенты женского пола имеют меньше несданных дисциплин;
2. Студенты ИШЯТ и ИШНКБ имеют меньше несданных дисциплин по сравнению с остальными;
3. Студенты, обучающиеся по целевому приему, имеют меньше несданных дисциплин, чем остальные;
4. Студенты, обучающиеся на специалитете имеют меньше несданных дисциплин по сравнению с остальными;
5. Студенты, имеющиеся гражданство Российской Федерации, имеют меньше несданных дисциплин;
6. Студенты-четверокурсники обучаются успешнее остальных;
7. Студенты, не состоящие в академическом отпуске, обучаются успешнее.

Заключение

Таким образом был предобработан набор данных о характеристиках студентов, проведен разведочный анализ, а также сформулированы и проверены статистические гипотезы. Результаты данной работы были использованы В.А. Галлингером при создании предсказательных моделей.

Список использованных источников

1. Губин Е.И. Методика подготовки больших данных для прогнозного анализа. / «Наука и бизнес: пути развития». Выпуск № 3(105). 2020, 2020. – [С. 33-35].
2. Губин Е.И. Методология подготовки больших данных для прогнозного анализа / Современные технологии, экономика и образование: Сборник трудов Всероссийской научно-методической конференции. / Томский политехнический университет. ¬– Томск: Изд-во Томского политехнического университета, 2019. – 139с. ¬– [С. 25-28].
3. Python для анализа данных. [Электронный ресурс]. – Режим доступа: https://www.coursera.org/specializations/machine-learning-data-analysis, свободный. – Загл. с экрана. (Дата обращения 27.02.2021).