СЕМЕРИКОВ А. В., ГЛАЗЫРИН М. А. ПЕРВИЧНЫЙ АНАЛИЗ СТРУКТУРЫ DATAFRAME СТУДЕНТОВ УНИВЕРСИТЕТА

УДК 378.141.21:330.47, ВАК 05.13.01:08.00.05, ГРНТИ 06.35.51

Первичный анализ структуры DataFrame студентов университета

А. В. Семериков[1], М. А. Глазырин[2]

[1]Ухтинский государственный технический университет, г. Ухта
[2]Вятский государственный университет, г. Киров

В статье представлен первичный анализ данных с использованием Pandas (библиотека Python). С помощью структуры DataFrame рассмотрен файл формата xlsx, в котором содержится обезличенное описание 37609 студентов по 13 признакам: институт, специальность, форма обучения, категория, средний балл, пол, общежитие, семейное положение, медаль, тип школы, лет после школы, страна, город. В качестве целевого признак принят факт окончания. При выполнении всей учебной программы этот признак имеет значение 1, в противном случае он равен 0. Также он может принимать значение от 0 до 100. Первичный анализ данных выявил характерные особенности рассматриваемых данных, которые в процессе машинного обучения могут не позволить построить деревья принятия решения не большой глубины, при кластеризации представляется проблемным снизить размерность главных компонент.

Ключевые слова: *большие данные, объекты, признаки, целевой признак, Pandas, библиотека Python.*

Simulation of a process model of functioning of the enterprises for rendering of services

A.V. Semerikov[1], M.A. Glazyrin[2]

[1]Ukhta State Technical University, Ukhta
[2]Vyatka State University, Kirov

This article introduces primary data analysis using Pandas (Python library). Using the DataFrame structure, an xlsx file is considered, which contains a description of 37609 students by 13 characteristics: institute, specialty, form of study, category, GPA, gender, dormitory, marital status, medal, type of school, years after school, country, city. The fact of completion is taken as the target feature. During the execution of the entire curriculum, this feature has a value of 1, otherwise it is equal to 0. It can also take a value from 0 to 100. The primary data analysis revealed the characteristic features of the data under consideration, which in the process of machine learning may not allow building decision trees. large depth, during clustering it seems problematic to reduce the dimension of the principal components.

Keywords: *big data, objects, features, target trait, Pandas, Python library.*

Введение

После окончания обучения в среднем учебном заведении многие школьники имеют желание продолжить обучение в высших учебных заведениях (ВУЗ). Для поступления в ВУЗ они предоставляют персональные данные, которые после зачисления расширяются и уточняются.

Набор данных о каждом студенте состоит из параметров: год поступления, название института, специальность, форма обучения, категория конкурса, сумма баллов, средняя сумма баллов, инвалидность, льготы, должность, страна, регион, город, процент окончания.

Последний параметр представляет собой целевой признак. Если студент окончил ВУЗ, этот параметр имеет значение 100. В противном случае он имеет значение пропорциональное количеству закрытых сессий или дисциплин. Если, например, студент закрыл половину сессии, то целевой признак имеет значение 50. В таком виде этот целевой признак можно использовать для построения регрессионной функции при проведении машинного обучения с учителем. Для проведения машинного обучения с целью классификации студентов этот целевой признак заменяется. В этом случае, если студент успешно справился с программой университета, значению целевого признака присваивается значение 1, в противном случае его значение равно 0.

Имея большое количество данных о студентах, представляется возможным, используя методы машинного обучения, построить дерево решений, регрессионную функцию и осуществить процесс кластеризации. На основе дерева решений, регрессии и кластеризации можно предсказать будущее студента первокурсника, то есть определить наиболее вероятный исход. Закончит он обучение в университете с целевым признаком 100 или с каким-либо другим.

В результате машинного обучения [1] необходимо построить алгоритм, который с большой достоверностью определял бы значение целевого показателя у вновь прибывшего студента. При этом можно ожидать построения недообученного или переобученного алгоритма. В последнем случае созданный алгоритм хорошо определяет положение студента на текущих данных и плохо справляется с вновь прибывшими студентами. Это происходит из-за большого влияния первичных данных на значение целевого признака. В первом же случае наоборот алгоритм был получен при недостаточном количестве переборов (кроссвалидация) исходных данных. Поэтому, прежде чем заняться машинным обучением, необходимо провести первичный анализ данных, который позволит установить шумы и противоречия в первичных данных. Кроме того в ходе первичного анализа данных можно сделать предварительный прогноз целевого признака.

Экспериментальная часть

В настоящей статье представлен первичный анализ данных с использованием Pandas – это библиотека Python, предоставляющая широкие возможности для анализа данных. С ее помощью очень удобно загружать, обрабатывать и анализировать табличные данные с помощью SQL-подобных запросов. В связке с библиотеками Matplotlib и Seaborn появляется возможность удобного визуального анализа табличных данных [2].

В процессе обучения студентов в университете накопился большой объем данных, которые можно извлечь в виде таблицы Microsoft Excel в формате xlsx. Файл такого формата очень хорошо считывается с помощью библиотеки Pandas. Основными структурами в Pandas являются классы Series и DataFrame. Последний используется для представления данных, в которых строки соответствуют набору признаков для описания отдельного студента, а столбцы соответствуют этим признакам.

В настоящей статье представлены результаты анализа данных, относящихся к описанию отдельного студента по 13 признакам: институт, специальность, форма обучения, категория, средний балл, пол, общежитие, семейное положение, медаль, тип школы, лет после школы, страна, город. Целевой признак имеет название «Факт окончания» (Таблица 1).

Таблица 1. Исходный набор данных по каждому студенту

Институт	Специальность	Форма обучения	Категория	Средний балл	Пол	Общежитие	Семейное положение	Медаль	Тип школы	Лет после школы	Страна	Город	Факт окончания
СТИ	Электропривод и автоматика промышленных установок и технологических комплексов	очная	общий конкурс	63,3	м	да	не женат	нет медали	школа	5	Россия	Ухта	1
ИнЭУиИТ	Автоматизированные системы обработки информации и управления	заочная	общий конкурс	65	м	да	не женат	серебряная	училище	7	Россия	Ухта	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Для проведения исследования представленных данных с помощью библиотек Python, была проведена замена текстовых данных на числовые следующим образом. По каждой колонке определялось количество уникальных признаков. Затем каждому уникальному текстовому признаку присваивалось соответствующее число. Например, т. к. количество уникальных форм обучения два (очная и заочная), то в колонке форма обучения проставляются соответственно 0 и 1. Таким образом переходим к следующему набору данных (Таблица 2).

Таблица 2. Числовой набор данных по каждому студенту

Институт	Специальность	Форма обучения	Категория	Средний балл	Пол	Общежитие	Семейное положение	Медаль	Тип школы	Лет после школы	Страна	Город	Факт окончания
3	45	2	1	63,3		0	1	1	4	5	1	4	1
2	43	1	1	65		да	1	3	5	7	1	4	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Для проведения исследования представленных наборов данных представляется удобным использование интерактивной вычислительной среды Jupyter Notebook [1], с помощью которой осуществляется импорт библиотеки Pandas, которая позволяет прочитать исходную Таблицу 2 в формате Microsoft Excel (Рисунок 1).

	Институт	Специальность	Форма обучения	Категория	Средний балл	Пол	Общежитие	сем. Положение	Медаль	Тип школы	Лет после школ	Стра
0	1	1	1	1	52.7	1	0	1	1	1	1	
1	2	2	1	2	26.7	1	1	1	1	2	2	
2	3	3	1	1	63.3	1	0	1	1	2	1	
3	1	4	1	2	26.3	1	0	1	1	3	3	
4	3	5	1	1	44.7	1	0	1	1	1	4	

Рисунок 1. Числовой набор данных по каждому студенту в представлении Pandas

Расчет основных статистических показателей данных Таблицы 2, представленных в Таблице 4, показал, что при общем количестве обучавшихся студентов 37609 человек средний балл при поступлении в университет составляет 52,67. Из 37609 студентов 17133 студентов получили дипломы об образовании. Это составляет 46 % от общей численности студентов.

Таблица 4. Основные статистические показатели

Форма обучения	Категория	Средний балл	Пол	Общежитие	Сем. положение	Медаль	Тип школы	Лет после школы	Страна	Город	Факт окончания
37609	37609	37609	37609	37609	37609	37609	37609	37609	37609	37609	37609
1.76	1.25	52.67	0.66	0.43	1.81	1.05	4.64	12.19	1.09	92.80	0.46

При этом, согласно критерия согласия Пирсона и теста Шапиро-Уилк гипотеза о нормальности функции распределения признака «средний балл» отклоняется. Отклонение от нормального закона распределения свидетельствует о большом количестве студентов со средним баллом ниже 52,67. В этом можно убедится, если посмотреть на количество студентов по признаку «средний балл»: 52,7 балла у 7031 человек, 50 баллов у 529 человек, 52 балла у 479 человек и т.д.

Подсчет количества студентов по каждому институту показывает крайнюю неравномерность (Таблица 5).

Таблица 5. Количество студентов в институтах

Номер института	1	2	3	4	5	6	7	8	9
Количество студентов	12158	7380	8540	8751	166	176	1	12	1

Институты с номерами 7, 8, 9 можно отнести к шумам. Поэтому, перед проведением машинного обучения, записи строки с этими объектами удаляются. Тем самым качество обучения при построении дерева решения и проведении кластеризации должно улучшиться.

В представленных для анализа данных наблюдается слабая зависимость статистик (средних значений) признаков объектов от целевого признака. В этом можно убедится, сравнивая значения в Таблице 6.

Таблица 6. Значение средних признаков в зависимости от целевого признака

Институт	2,45	2,55
Специальность	10,03	42,05
Форма обучения	1,77	1,74
Категория	1,22	1,29
Средний балл	52,06	53,4
Пол	0,67	0,65
Общежитие	0,48	0,37
Семейное положение	1,73	1,93
Медаль	1,05	1,06
Тип школы	4,68	4,58
Лет после школы	12,09	12,3
Страна	1,05	1,33
Город	85,77	101,21
Факт окончания	0	1

Результаты

Анализ признаков объектов (студентов) позволяет сделать такие выводы:

1. Предварительный прогноз по целевому признаку. Примерно половина, обучающихся студентов, закончат институты с получением диплома (Таблица 7).

2. Высота дерева решения при этих данных будет большой.

3. В формуле регрессии все признаки будут иметь примерно одинаковые значения.

Таблица 7. Количество студентов в зависимости от факта окончания

Факт окончания	Институт											
	1	2	3	4	5	6	7	8	9	10	11	Все
0 (не закончил)	6454	4533	4576	4477	152	127	1	12	1	106	37	20476
1 (закончил)	5704	2847	3965	4274	14	49	0	0	0	89	191	17133
Все	12158	7380	8541	8751	166	176	1	12	1	195	228	37609

4. При проведении кластеризации не удастся выделить ограниченное количество главных компонент (снизить размерность до 2 или 3).

5. Студенты со средним количеством баллов больше 50 могут расположится в таком соотношении (Таблица 8).

6. Точность прогноза по целевому признаку ожидается не высокой

Таблица 8. Количество студентов в зависимости от факта окончания

Средний балл больше 50	Факт окончания		
	0 (не закончил)	1 (закончил)	Все
0 (нет)	7488	5668	13156
1 (да)	12988	11465	24453
Все	20476	17133	37609

На графике это выглядит так (Рисунок 2).

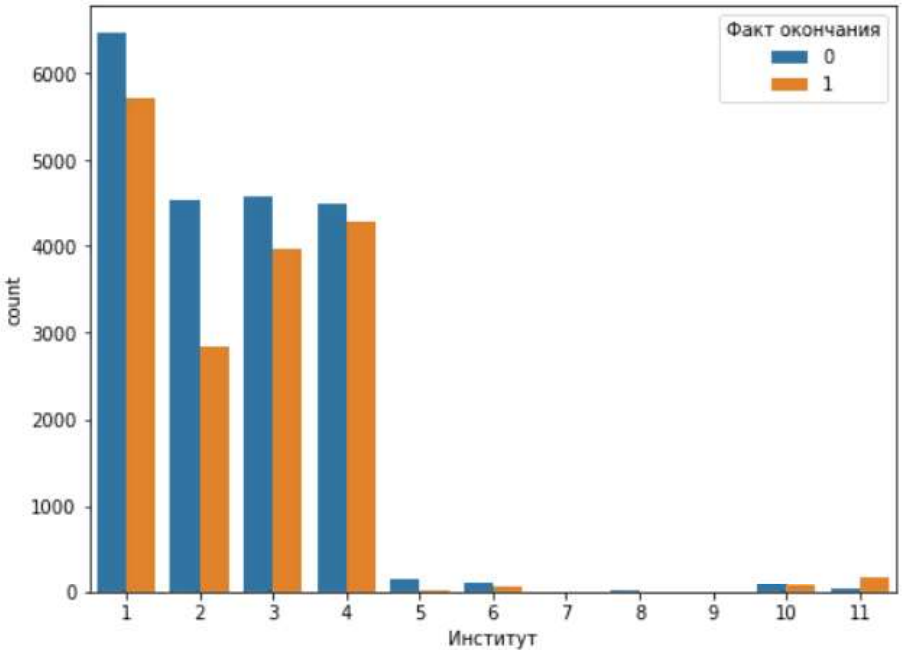

Рисунок 2. Количество студентов по институтам

Список использованных источников и литературы

1. Открытый курс по машинному обучению [Электронный ресурс]. – Режим доступа: https://www.youtube.com/watch?v=p9Hny3Cs6rk (дата обращения: 11.02.2021).

List of references

1. Open Course in Machine Learning, https://www.youtube.com/watch?v=p9Hny3Cs6rk, accessed Febrary 11, 2021.