Отчёт об анализе данных отсутствия на работе

Динара Руслановна Файзуллина

2 ноября 2018 г.

Описание данных

Анализировались данные об отсутствии сотрудников курьерной фирмы в Бразилии на рабочем месте, которые собирались с июля 2007 года по июль 2010 года.

Данные содержат 740 записей и 21 переменную.

Описательные статистики

Характеристики метрических переменных исследуемого набора данных представлены в таблице:

Переменная	Минимум	Среднее	Медиана	Максимум
Transportation.expense	118.0	221.3	225.0	388.0
Distance.from.Residence.to.Work	5.00	29.63	26.00	52.00
Service.time	1.00	12.55	13.00	29.00
Age	27.00	36.45	37.00	58.00
Work.load.Average.day	205.9	271.5	264.2	378.9
Son	0.000	1.019	1.000	4.000
Pet	0.0000	0.7459	0.0000	8.0000
Weight	56.00	79.04	83.00	108.00
Height	163.0	172.1	170.0	196.0
Body.mass.index	19.00	26.68	25.00	38.00
Absenteeism.time.in.hours	0.000	6.924	3.000	120.000

Так, Pet изменяется от 0 до 8 со средним значением равным 0.7459 и медианой равной 0, что говорит о наличии выбросов в данных.

Аналогично, Absenteeism.time.in.hours изменяется от 0 до 120, тогда как среднее значение равно 6.924, а медиана равна 3. Откуда следует, что значение данной

переменной является небольшим для большинства записей, так как даже выброс в виде максимального значения не сильно повлиял на среднее.

Для описания набора данных также использовались следующие переменные:

1. Идентификационная переменная – id сотрудника (ID): от 1 до 36
2. Переменная категории – причина отсутствия (Reason.for.absence)[1]: от 1 до 28
3. Переменная категории – месяц (Month.of.absence): от 1 до 12
4. Переменная категории – день недели (Day.of.the.week)[2]: от 2 до до 6
5. Переменная категории – время года (Seasons): от 1 до 4
6. Бинарная переменная – дисциплинарное взыскание (Disciplinary.failure)
7. Переменная категории – образование (Education)[3]
8. Бинарная переменная – употребление алкоголя (Social.drinker)
9. Бинарная переменная – курение (Social.smoker)

Распределение выборки по показателю Month.of.absence:
1 – 6.7%, 2 – 9.7%, 3 – 11.7%, 4 – 7.16%, 5 – 8.6%, 6 – 7.3%, 7 – 9%, 8 – 7.3%, 9 – 7.2%, 10 – 9.6%, 11 – 8.5%, 12 – 6.6%

Распределение выборки по показателю Day.of.the.week:
2 – 21.7%, 3 – 20.8%, 4 – 21.1%, 5 – 16.9%, 6 – 19.5%

Распределение выборки по показателю Seasons:
1 – 22.9%, 2 –25.9%, 3 – 24.7%, 4 – 26.3%

Распределение выборки по показателю Education:
1 – 82.5%, 2 – 6.2%, 3 – 10.6%, 4 – 0.5%

Распределение выборки по показателю Reason.for.absence:
1 – 2.1%, 2 – 0.1%, 3 – 0.1%, 4 – 0.2%, 5 – 0.4%, 6 – 1.0%, 7 – 2.0%, 8 – 0.8%, 9 – 0.5%, 10 – 3.3%, 11 – 3.5%, 12 – 1.0%, 13 – 7.4%, 14 – 2.5%, 15 – 0,2%, 16 – 0.4%, 17 – 0.1%, 18 – 2.8%, 19 – 5.4%, 20 – 0%, 21 – 0.8%, 22 – 5.1%, 23 – 20.1%, 24 – 0.4%, 25 – 4.1%, 26 – 4.4%, 27 – 9.3%, 28 – 15.1%

[1]Причины отсутствия 1-21 были засвидетельствованы Международной классификацией болезней (МСБ), причины 22-28 без МСБ

[2]Понедельник (2), Вторник (3), Среда (4), Четверг (5), Пятница (6)

[3]High school (1), graduate (2), postgraduate (3), master and doctor (4)

T-тест

Рассмотрим различия в уровне Absenteeism.time.in.hours по показателям Social.driker и Social.smoker. Согласно критерию Стьюдента (t-test) не выявлены статистически значимые различия между группой 0 и группой 1 (p = 0.07396, t = -1.7895 и p = 0.756, t = 0.31204 соответственно).

Однако по показателю Disciplinary.failure нам удалось выявить статистически значимые различия между группой 0 и группой 1 в уровне Absenteeism.time.in.hours (p < 2.2e-16, t = 14.239, средние значения 7.32 и 0.00 соответственно).

Интересно заметить, что по показателю Social.drinker были выявлены различия между группами 0 и 1 в уровне переменных, перечисленных ниже. Диаграмма размаха для переменной Weight представлена на Рис.1.

- Distance.from.Residence.to.Work (p < 2.2e-16, t = -14.672, means 21.95 & 35.48)
- Service.time (p < 2.2e-16, t = -10.443, means 10.78 & 13.90)
- Weight (p < 2.2e-16, t = -11.17, means 73.45 & 83.29)
- Body.mass.index (p < 2.2e-16, t = -9.3024, means 25.08 & 27.09)
- Age (p = 2.305e-08, t = -5.6702, means 34.86 & 37.65)
- Son (p = 1.926e-09, t = -6.0829, means 0.75 & 1.21)

В то время как по показателю Social.smoker наблюдаются различия между группами 0 и 1 по следующим переменным:

- Age (p = 0.0001528, t = -4.0159, means 36.22 & 39.25)
- Son (p = 2.192e-08, t = -6.2567, means 0.97 & 1.62)
- Weight (p = 1.16e-11, t = 8.0042, means 79.75 & 69.92)
- Body.mass.index (p = 2.483e-08, t = 6.3338, means 26.91 & 23.68)

Хи-квадрат

Для выявления взаимосвязей между номинальными переменными используется критерий хи-квадрат.

Между номинальными переменными, перечисленными ниже, было показано наличие статистически значимой взаимосвязи:

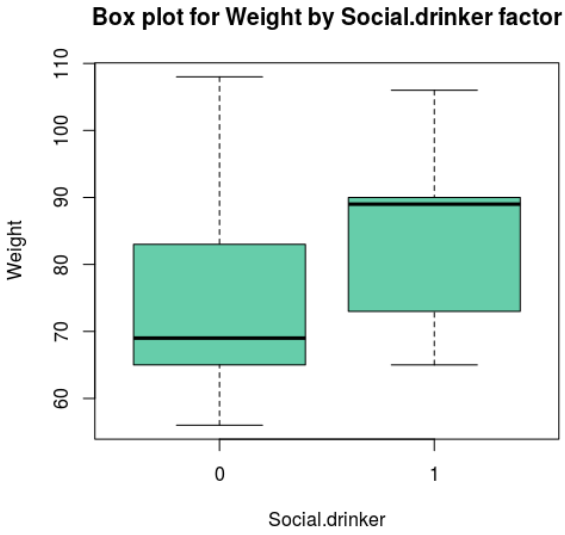

Рис. 1: Диаграмма "ящик с усами"для Social.drinker и Weight

- Reason.for.absence и Month.of.absence (p<2.2e-16, X-squared = 599.52, df = 324)
- Reason.for.absence и Seasons (p<2.2e-16, X-squared = 267.86, df = 81)
- Reason.for.absence и Education(p = 1.237e-10, X-squared = 189.13, df = 81)
- Reason.for.absence и Disciplinary.failure (p<2.2e-16, X-squared = 685, df = 27)
- Reason.for.absence и Social.drinker (p = 1.757e-08, X-squared = 88.612, df = 27)
- Reason.for.absence и Social.smoker (p = 2.766e-09, X-squared = 93.654, df = 27)

ANOVA

Рассмотрим различия в уровне различных переменных по показателю Reason.for.absence, разделяющему выборку на 28 групп. Согласно анализу ANOVA были выявлены статистически значимые различия между группами в уровне Absenteeism.time.in.hours (p = 2.17e-06, F = 22.8, signif.level = 0.001), Distance.from. Residence.to.Work (p = 9.69e-06, F = 19.85, signif.level = 0.001), Work.load.Average.day (p = 0.000763, F = 11.43, signif.level = 0.001).

По показателю Education, разделяющему выборку на 4 группы, были также выявлены статистически значимые различия в уровне переменных: Service.time(p = 4.87e-09, F = 35.07, signif.level = 0.001), Age (p = 1.05e-09, F = 38.21, signif.level = 0.001), Son (p = 2.36e-07, F = 27.23, signif.level = 0.001), Weight (p <2e-16, F = 73.3, signif.level = 0.001), Body.mass.index (p <2e-16, F = 73.3, signif.level = 0.001).

Корреляционный анализ

Корреляционный анализ позволяет определить взаимосвязь между метрическими переменными. Значения коэффициентов корреляции представлены в таблице, статистически значимые взаимосвязи выделены полужирным шрифтом.[4]

	Transport	Distance	Service	Age	Work	Son	Pet	Weight	Height	Index	Absent
Transport	1.00	**0.26**	**-0.35**	**-0.23**	0.01	**0.38**	**0.40**	**-0.21**	**-0.19**	**-0.14**	0.03
Distance	**0.26**	1.00	**0.13**	**-0.15**	-0.07	0.05	**0.21**	-0.05	**-0.35**	**0.11**	**-0.09**
Service	**-0.35**	**0.13**	1.00	**0.67**	-0.00	-0.05	**-0.44**	**0.46**	-0.05	**0.50**	0.02
Age	**-0.23**	**-0.15**	0.67	1.00	-0.04	0.06	**-0.23**	**0.42**	-0.06	**0.47**	0.07
Work	0.01	-0.07	-0.00	-0.04	1.00	0.03	0.01	-0.04	**0.10**	**-0.09**	0.02
Son	**0.38**	0.05	-0.05	0.06	0.03	1.00	**0.11**	**-0.14**	-0.01	**-0.14**	**0.11**
Pet	**0.40**	**0.21**	**-0.44**	**-0.23**	0.01	**0.11**	1.00	**-0.10**	**-0.10**	**-0.08**	-0.03
Weight	**-0.21**	-0.05	**0.46**	**0.42**	-0.04	**-0.14**	**-0.10**	1.00	**0.31**	**0.90**	0.02
Height	**-0.19**	**-0.35**	-0.05	-0.06	**0.10**	-0.01	**-0.10**	**0.31**	1.00	**-0.12**	**0.14**
Index	**-0.14**	**0.11**	**0.50**	**0.47**	**-0.09**	**-0.14**	**-0.08**	**0.90**	**-0.12**	1.00	-0.05
Absent	0.03	**-0.09**	0.02	0.07	0.02	**0.11**	-0.03	0.02	**0.14**	-0.05	1.00

В частности, Service.time прямо связана с Age, причем связь достаточно сильная ($r = 0.6709789$, $p < 2.2e\text{-}16$). Аналогично, Service.time связано достаточно сильно с Body.mass.index ($r = 0.499718$, $p < 2.2e\text{-}16$).

Очевидно, что Body.mass.index очень сильно связана с Weight ($p = 0.9041169$, $p < 2.2e\text{-}16$).

Тогда как связь между Absenteeism.time.in.hours с Transportation.expense, Distance.from.Residence.to.Work, Son оказалась достаточно слабой ($r = 0.02758463$, $p = 0.4537$; $r = -0.08836282$, $p = 0.0162$; $r = 0.1137565$, $p = 0.001939$).

Корреляционная плеяда

Результаты корреляционного анализа можно визуализировать в виде корреляционной плеяды, представленной на Рис. 2.

Регрессионный анализ

Регрессионный анализ позволяет исследовать влияние одной или нескольких независимых переменных $x_1, x_2, \ldots, x_n$ на зависимую переменную y.

В качестве интересных для исследования зависимых переменных были взяты Absenteeism.time.in.hours и Service.time. Методом последовательного удаления из

[4] Имена переменных были сокращены с сохранением семантики, а размер шрифта уменьшен с целью лучшей визуализации статистических данных

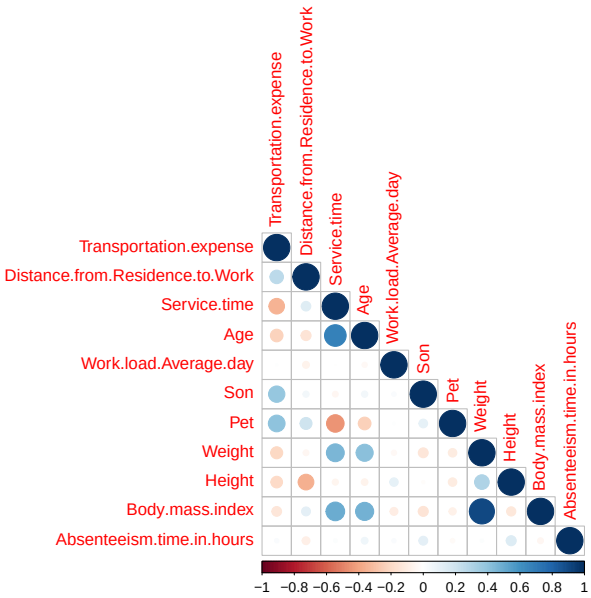

Рис. 2: Корреляционная плеяда для набора данных Absenteeism_at_work

списка зависимых переменных тех, что имеют с высокое значение p-value, и следовательно не влияют на исследуемую зависимую переменную, были сформированы модели.

Для Absenteeism.time.in.hours не удалось подобрать модель, которая бы описывала больше 5% данных, поэтому рассмотрим наилучшую модель для Service.time.

Для оценки результатов регрессии с помощью метода кросс-валидации исходный набор данных был разбит на тренировочный (80%) и тестовый (20%) наборы. На основе тренировочного набора было проведен регрессионный анализ.

Определим зависимость между Service.time(y) и такими переменными, как Transportation.expense(x_1), Distance.from.Residence.to.Work(x_2), Age(x_3), Pet(x_4), Weight(x_5), Height(x_6), Body.mass.index(x_7). Значения коэффициентов регрессионного уравнения и уровни значимости представлены в таблице ниже.

	Estimate	Std.Error	t-value	p-value	Signif. level
(Intercept)	73.409211	15.564845	4.716	3.01e-06	0.001
Transportation.expense	-0.009375	0.001763	-5.317	1.50e-07	0.001
Distance.from.Residence.to.Work	0.095464	0.008035	11.882	$<$ 2e-16	0.001
Age	0.368652	0.019447	18.957	$<$ 2e-16	0.001
Pet	-1.101687	0.086687	-12.709	$<$ 2e-16	0.001
Weight	0.568687	0.100164	5.678	2.15e-08	0.001
Height	-0.466953	0.091389	-5.110	4.38e-07	0.001
Body.mass.index	-1.452768	0.290937	-4.993	7.84e-07	0.001

Уравнение регрессии выглядит следующим образом:

$$y = -0.009375x_1 + 0.095464x_2 + 0.368652x_3 - 1.101687x_4 + 0.568687x_5 - 0.466953x_6 - 1.452768x_7 + 73.409211$$

Отметим, что данная модель объясняет 68% изменчивости данных (Multiple R-squared: 0.6857, Adjusted R-squared: 0.682)

Доверительные интервалы для каждого коэффициента уравнения с доверительной вероятностью 95%:

	2.5 %	97.5 %
(Intercept)	42.84	103.98
Transportation.expense	-0.01	-0.01
Distance.from.Residence.to.Work	0.08	0.11
Age	0.33	0.41
Pet	-1.27	-0.93
Weight	0.37	0.77
Height	-0.65	-0.29
Body.mass.index	-2.02	-0.88

Для оценки качества регрессии проведем корреляционный анализ для ожидамого и полученного значений переменной sales на тренировочном и тестовом наборах. Квадрат корреляции на тренировочном наборе равен 0.6889, что практически совпадает с Adjusted R-squared, равным 0.682. Квадрат корреляции на тестовом наборе равен 0.6889, что полностью совпадает Adjusted R-squared.

Следовательно, можно сделать вывод, что данное уравнение регрессии предсказывает значение зависимой переменной с высокой точностью.

Кластеризация

Процедура кластерного анализа позволяет упорядочить объекты выборки в сравнительно однородные группы на основе информации о наборе данных.

Применим один из наиболее популярных методов кластеризации — метод k-средних. Перед кластерным анализом данные также были нормализированы, чтобы среднее по каждой переменной было равным 0, а std error 1.

Для определения оптимального количества кластеров использовался elbow method. На графике каменистой диаграммы, изображенном на Рис. 3, видно, что число кластеров $k = 5$ можно принять за оптимальное, так как разница между суммами квадратов для большего числа кластеров изменяется не сильно.

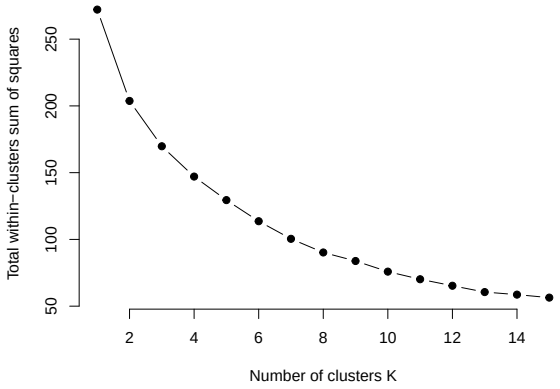

Рис. 3: Диаграмма суммы квадратов внутри кластеров для k.max = 15

После разделения набора данных на 5 кластеров было проведено построение clusplot, изображенное на Рис. 4, где каждый из кластеров обозначен уникальным цветом. Визуализация представлена по двум компонентам, по которым данные наиболее сильно изменяются.

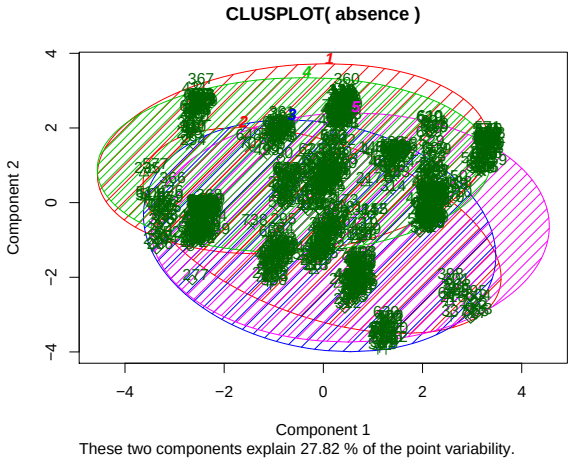

Рис. 4: Clusplot для двух компонент

В таблице ниже представлены средние значения переменных для каждого из 5 кластеров в исходных единицах, а не нормализованных:

№	Transport	Distance	Service	Age	Work	Son	Pet	Weight	Height	Index	Absent
1	139.95	12.77	16.34	42.68	280.49	1.51	0.12	95.80	185.63	28.14	6.52
2	200.34	42.61	16.89	40.60	263.87	0.46	0.45	91.10	168.81	32.16	4.09

3	204.29	19.93	13.07	39.93	266.94	1.57	0.43	79.21	175.71	25.57	88.00
4	189.77	18.68	11.64	36.31	272.17	0.41	0.14	71.16	171.92	23.99	5.19
5	278.51	34.49	9.66	32.35	273.80	1.69	1.56	73.70	170.95	25.07	6.00

На Рис. 5 изображено распределение данных на кластеры при сравнении значений переменных Absenteeism.time.in.hours и Transportation.expense на наборе данных.

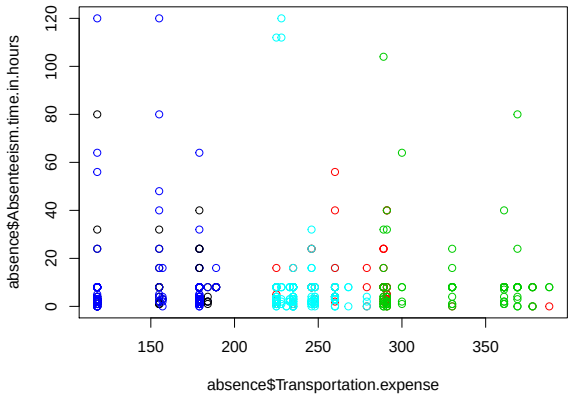

Рис. 5: Диаграмма рассеивания для характеристик Absenteeism.time.in.hours и Transportation.expense с выделенными кластерами