1

Lập trình cho khoa học dữ liệu

Thành viên:

- Nguyễn Đăng Quang – 18120527
- Nguyễn Văn Tuấn Đạt – 19120472
- Phan Xuân Hoài – 20120481
- Luân Mã Khương – 20120515

GVHD:

- Bùi Tiến Lên
- Lê Nhựt Nam
- Trần Đại Chí

2


Tổng kết
Thu thập dữ liệu
Đưa ra các câu hỏi có ý nghĩa cần trả lời
Tiền xử lý dữ liệu

3

Thu thập dữ liệu

4

Về gói dữ liệu

Một công ty về dữ liệu muốn thuê các Data Scientist và dữ liệu này được thu thập từ các ứng viên đăng ký vào công ty

Gói dữ liệu này là thông tin về các ứng viên bao gồm giới tính, trình độ học vấn, kinh nghiệm làm việc,...

Nguồn: https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

5

Thông tin về gói dữ liệu

Gói dữ liệu bao gồm:

- 19158 dòng: ứng với các ứng viên
- 14 cột: ứng với các thông tin thuộc về ứng viên

6

Thông tin về gói dữ liệu

- enrollee_id : ID với từng ứng viên
- city: Mã thành phố
- city_development_index : Chỉ số phát triển của thành phố (theo tỷ lệ)
- gender: Giới tính
- relevent_experience: Kinh nghiệm liên quan của ứng viên
- enrolled_university: Loại khóa học đại học đã đăng ký nếu có
- education_level: Trình độ học vấn
- major_discipline: Ngành học chính
- experience: Kinh nghiệm (tính theo năm)
- company_size: Số lượng nhân viên trong công ty của người sử dụng lao động hiện tại
- company_type : Loại chủ lao động hiện tại
- last_new_job: Sự khác biệt về số năm giữa công việc trước đây và công việc hiện tại
- training_hours: Thời gian hoàn thành huấn luyện
- target: 0 – Không nhảy việc, 1 – Nhảy việc

7

Khám phá dữ liệu

8

Kiểm tra dữ liệu trùng lặp

```
have_duplicated_rows = np.any(df.duplicated())
have_duplicated_rows
```

```
False
```

Các dòng không bị trùng lặp.

9

Không có dữ liệu trùng lặp.

Các cột có kiểu dữ liệu nào chưa phù hợp để tiếp tục sử lý hay không?

```
df.dtypes
```

```
enrollee_id                int64
city                      object
city_development_index    float64
gender                    object
relevent_experience       object
enrolled_university       object
education_level           object
major_discipline          object
experience                object
company_size              object
company_type              object
last_new_job              object
training_hours             int64
target                    float64
dtype: object
```

10

Vì cột experience có các dữ liệu như <1, 1, 2, ..., 20, >20 nên ta sẽ chuyển dữ liệu sang dạng các khoảng như (0,1), (1,5), (6,10), (11,15), (16,20), (21,30)

0	(21, 30)
1	(11, 15)
2	(1, 5)
3	(0, 1)
4	(21, 30)
	...
19153	(11, 15)
19154	(11, 15)
19155	(21, 30)
19156	(0, 1)
19157	(1, 5)

11

Các giá trị được phân bố như thế nào trong mỗi cột?

12

Các cột dữ liệu dạng số

Các cột dạng số: *'enrollee_id', 'city_development_index', 'training_hours', 'target'*

Thông tin phân bố giá trị của các cột sẽ được lưu vào Dataframe 'sumamary_df':

- 4 cột ứng với 4 cột dạng số trên
- 9 dòng:
  - ***missing_ratio:*** tỉ lệ phần trăm các giá trị thiếu
  - ***count:*** số lượng giá trị
  - ***mean:*** giá trị trung bình
  - ***std:*** độ lệch chuẩn
  - ***min:*** giá trị nhỏ nhất
  - ***25%:*** giá trị phân vị 25%
  - ***50%:*** giá trị phân vị 50%
  - ***75%:*** giá trị phân vị 75%
  - ***max:*** giá trị lớn nhất

13

Các cột dữ liệu dạng số

	enrollee_id	city_development_index	training_hours	target
missing_ratio	0.0	0.0	0.0	0.0
count	19158.0	19158.0	19158.0	19158.0
mean	16875.4	0.8	65.4	0.2
std	9616.3	0.1	60.1	0.4
min	1.0	0.4	1.0	0.0
25%	8554.2	0.7	23.0	0.0
50%	16982.5	0.9	47.0	0.0
75%	25169.8	0.9	88.0	0.0
max	33380.0	0.9	336.0	1.0

Như vậy các cột này không có giá trị thiếu

14

Cột dữ liệu categorical

Các giá trị sẽ tính:

- Tỉ lệ % (từ 0 đến 100) các giá trị thiếu
- Số lượng các giá trị khác nhau (không xét giá trị thiếu)
- Tỉ lệ % (từ 0 đến 100) của mỗi giá trị xếp theo tỉ lệ % giảm dần

	city	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type	last_new
missing_ratio	0.0	23.53064	0.0	2.014824	2.401086	14.683161	0.339284	30.994885	32.049274	2.207
num_values	123	3	2	3	5	6	6	8	6	
value_ratios	{'city_103': 22.7, 'city_21': 14.1, 'city_16':...	{'Male': 90.2, 'Female': 8.5, 'Other': 1.3}	{'Has relevent experience': 72.0, 'No relevent...	{'no_enrollment': 73.6, 'Full time course': 20...	{'Graduate': 62.0, 'Masters': 23.3, 'High Scho...	{'STEM': 88.7, 'Humanities': 4.1, 'Other': 2.3...	{(1, 5): 30.7, (6, 10): 26.2, (21, 30): 17.2, ...	{'50-99': 23.3, '100-500': 19.4, '10000+': 15....	{'Pvt Ltd': 75.4, 'Funded Startup': 7.7, 'Publ...	{'1': 42.9, 17.6, '2': 'never':

15

Tiền xử lý dữ liệu

16

Tiền xử lý dữ liệu

Giá trị bị thiếu

enrollee_id	0
city	0
city_development_index	0
gender	4508
relevent_experience	0
enrolled_university	386
education_level	460
major_discipline	2813
experience	65
company_size	5938
company_type	6140
last_new_job	423
training_hours	0
target	0

Ta thấy có khá nhiều giá trị bị thiếu và đa số đều là categorical. Vì vậy, ta sẽ xóa các dòng bị thiếu.

17

Tiền xử lý dữ liệu

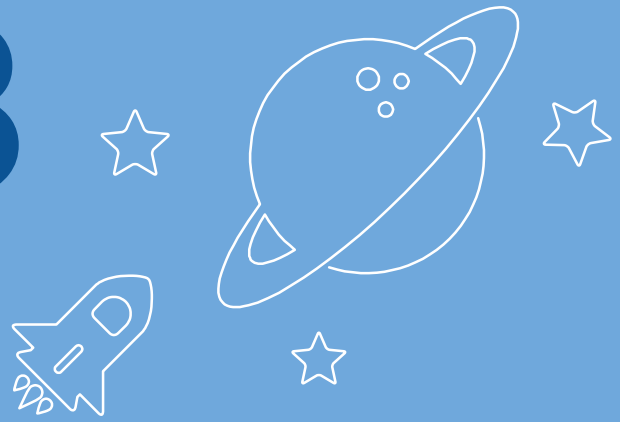
Giá trị bị thiếu

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_si
1	29725	city_40	0.776	Male	No relevent experience	no_enrollment	Graduate	STEM	(11, 15)	50-
4	666	city_162	0.767	Male	Has relevent experience	no_enrollment	Masters	STEM	(21, 30)	50-
7	402	city_46	0.762	Male	Has relevent experience	no_enrollment	Graduate	STEM	(11, 15)	<
8	27107	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	(6, 10)	50-
11	23853	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	(1, 5)	5000-99
...	...	...	...	...	...	...	...	...	...	
19147	21319	city_21	0.624	Male	No relevent experience	Full time course	Graduate	STEM	(1, 5)	100-5
19149	251	city_103	0.920	Male	Has relevent experience	no_enrollment	Masters	STEM	(6, 10)	50-
19150	32313	city_160	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM	(6, 10)	100-5
19152	29754	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	Humanities	(6, 10)	10/
19155	24576	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	(21, 30)	50-

9518 rows × 14 columns

Sau khi xóa, dữ liệu còn 9518 dòng.
Số lượng giá trị thiếu chiếm khoảng 50% dữ liệu.

18



Đưa ra các câu hỏi có ý nghĩa cần trả lời

19

Đưa ra các câu hỏi có ý nghĩa cần trả lời

1. Tỉ lệ ứng viên vượt qua khóa học Data Scientist có trên 5 năm kinh nghiệm làm việc đang tìm kiếm công việc Data Scientist trong Top 5 thành phố phát triển nhất?
2. Số năm kinh nghiệm và kinh nghiệm có liên quan đến DS ảnh hưởng thế nào đến quyết định có nhảy việc không của người tham gia đào tạo?
3. Có sự chênh lệch trình độ học vấn giữa nam và nữ hay không? Điều này có ảnh hưởng tới chỉ số phát triển của thành phố nơi mà họ làm việc?
4. Trình độ học vấn và kinh nghiệm làm việc của ứng viên tại các công ty lớn?
5. Sau khi hoàn thành huấn luyện ở các công ty nhỏ, ứng viên có quyết định nhảy việc hay không?

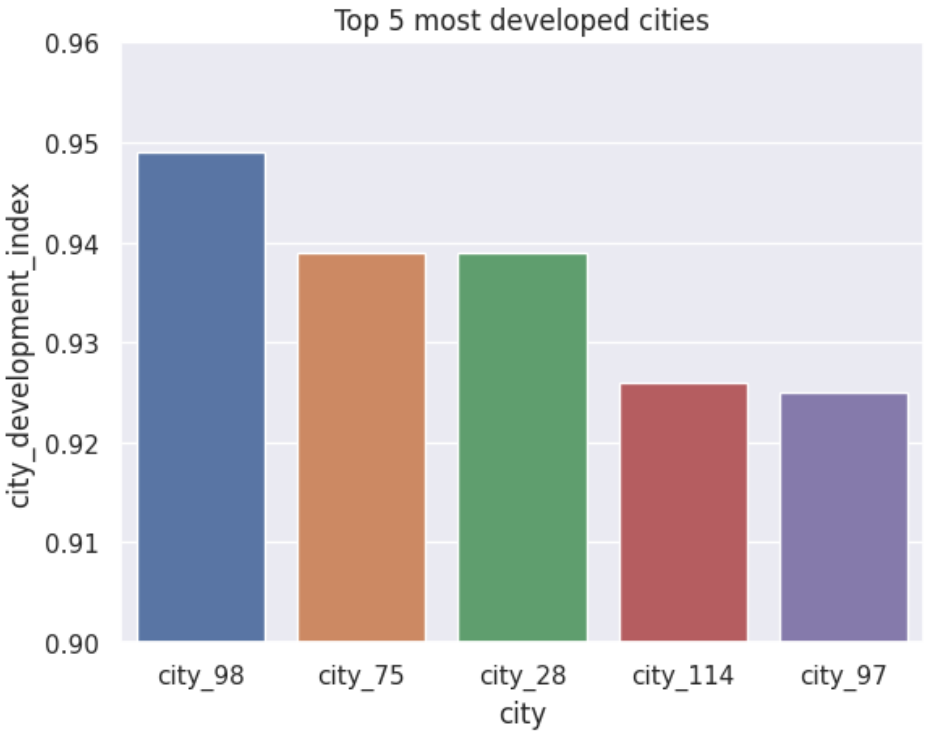
20

Đưa ra các câu hỏi có ý nghĩa cần trả lời

Tỉ lệ ứng viên vượt qua khóa học Data Scientist có trên 5 năm kinh nghiệm làm việc đang tìm kiếm công việc Data Scientist trong Top 5 thành phố phát triển nhất?

- Với vai trò là doanh nghiệp sẽ biết thêm thông tin để thay đổi chính sách đãi ngộ hay điều chỉnh môi trường làm việc để thu hút các ứng viên nhảy việc.
- Liệu các nhân viên nhiều năm kinh nghiệm có nhu cầu tìm kiếm thách thức mới cho bản thân mình.

Bước 1: Tìm Top 5 thành phố phát triển nhất dựa trên "city_development_index" và gán vào biến top_5_city

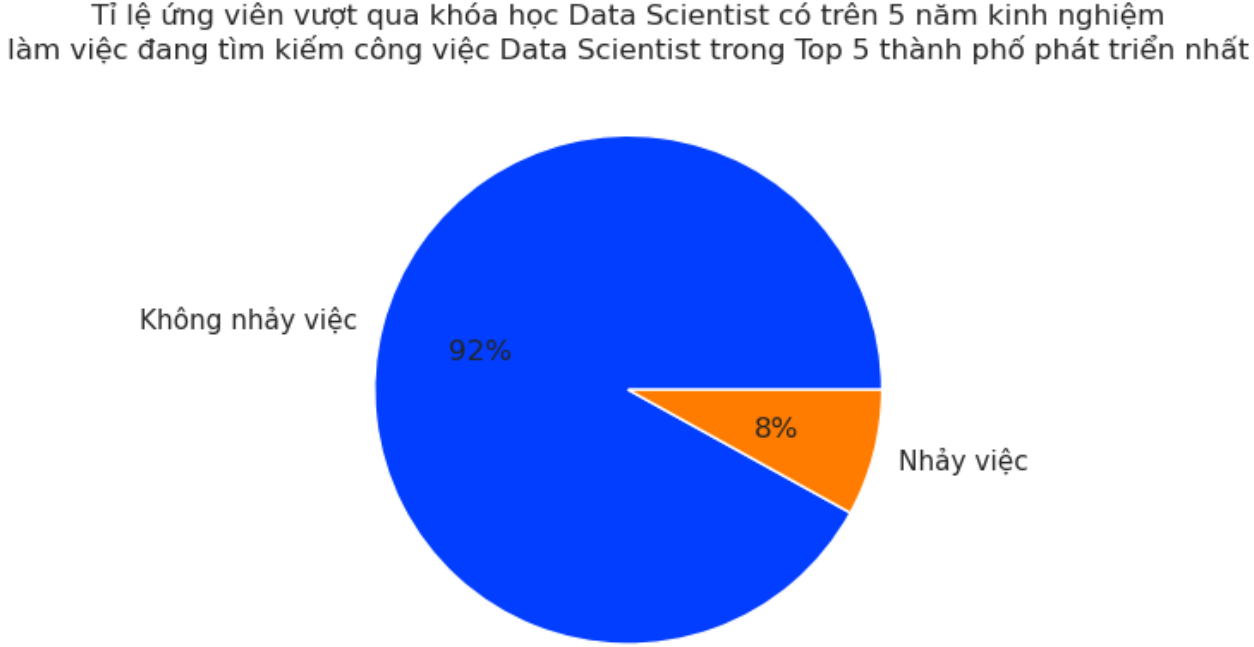


Bước 2: Tìm ứng viên vượt qua khóa học Data Scientist có trên 5 năm kinh nghiệm làm việc trong Top 5 thành phố phát triển nhất và gán vào biến DS_developed_city

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_si
15	6588	city_114	0.926	Male	Has relevent experience	no_enrollment	Graduate	STEM	(16, 20)	Oct-
21	19061	city_114	0.926	Male	Has relevent experience	no_enrollment	Masters	STEM	(11, 15)	100-5
37	10164	city_114	0.926	Male	Has relevent experience	no_enrollment	Phd	STEM	(21, 30)	100-5
40	2547	city_114	0.926	Female	Has relevent experience	Full time course	Masters	STEM	(16, 20)	1000-49
61	26516	city_75	0.939	Male	Has relevent experience	no_enrollment	Graduate	STEM	(6, 10)	100-5
...	...	...	...	...	...	...	...	...	...	
19029	11482	city_114	0.926	Male	Has relevent experience	no_enrollment	Masters	STEM	(11, 15)	50-
19064	12193	city_97	0.925	Male	Has relevent experience	no_enrollment	Masters	STEM	(21, 30)	1000
19071	17005	city_75	0.939	Male	No relevent experience	no_enrollment	Graduate	No Major	(21, 30)	1000-49
19077	16094	city_114	0.926	Male	Has relevent experience	no_enrollment	Masters	STEM	(11, 15)	50-
19089	11669	city_114	0.926	Male	Has relevent experience	Full time course	Masters	STEM	(16, 20)	50-

896 rows × 14 columns

Bước 3: Tính tỉ lệ các ứng viên vượt qua khóa học Data Scientist có trên 5 năm kinh nghiệm làm việc đang tìm kiếm công việc Data Scientist trong Top 5 thành phố phát triển nhất?



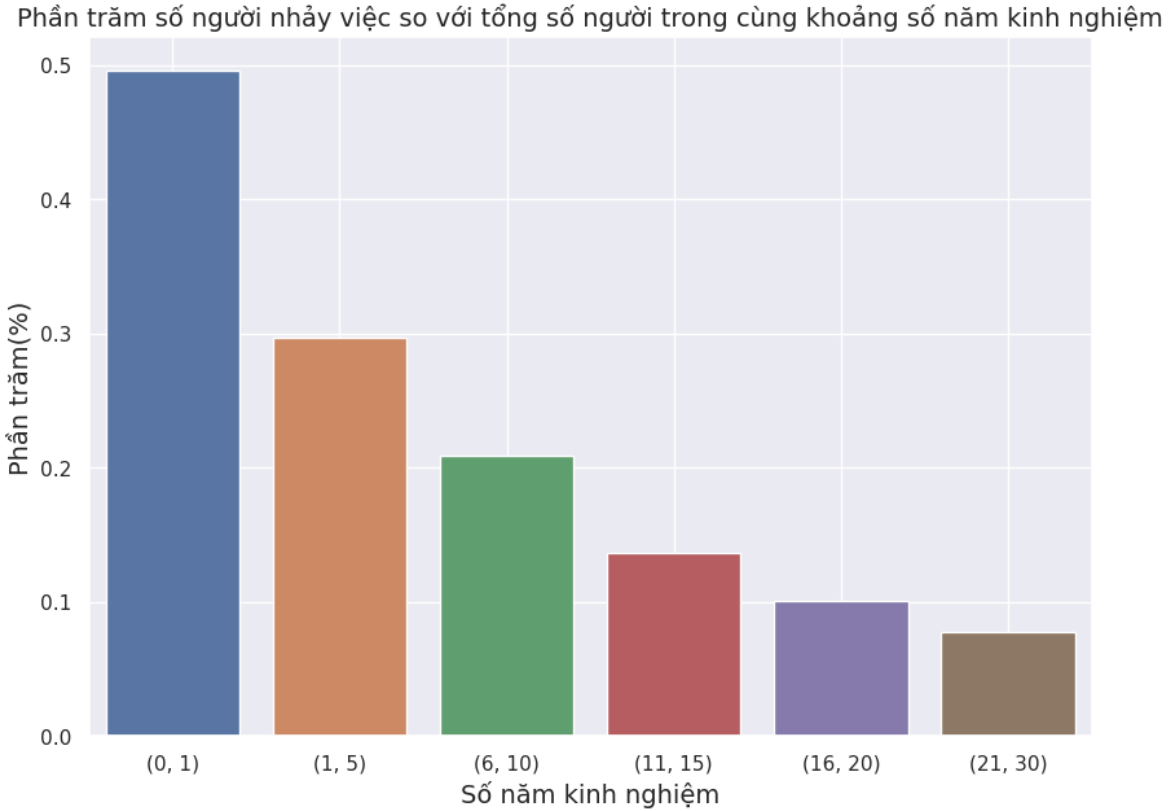
24

Đưa ra các câu hỏi có ý nghĩa cần trả lời

Số năm kinh nghiệm và kinh nghiệm có liên quan đến DS ảnh hưởng thế nào đến quyết định có nhảy việc không của người tham gia đào tạo?

- Công ty có thể đánh giá được đâu là một ứng viên tiềm năng muốn làm việc cho công ty dựa vào số năm kinh nghiệm và kinh nghiệm có liên quan đến DS của ứng viên để

Bước 1: Tương quan giữa số năm kinh nghiệm và quyết định nhảy việc



Bước 1: Tương quan giữa kinh nghiệm liên quan đến DS và quyết định nhảy việc

27

Đưa ra các câu hỏi có ý nghĩa cần trả lời

Có sự chênh lệch trình độ học vấn giữa nam và nữ hay không? Điều này có ảnh hưởng tới chỉ số phát triển của thành phố nơi m

- Thống kê được sự phân bố của các ứng viên tới các thành phố có chỉ số phát triển cao hay thấp liên quan tới trình độ học vấn của các ứng viên phân nhóm theo giới tínhà họ làm việc?

Bước 1: Thống kê số lượng các ứng viên theo giới tính

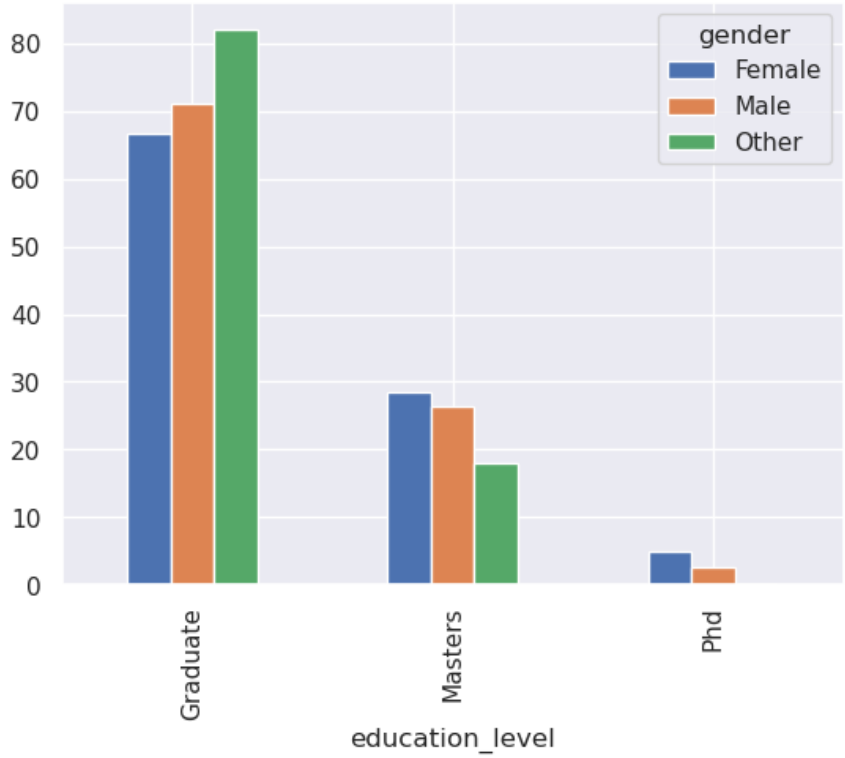
```
gender
Female     857
Male      8572
Other       89
dtype: int64
```

Ta thấy số lượng ứng viên nam có 1 sự chênh lệch rõ rệt so với số lượng ứng viên với các giới tính khác

Bước 2: Thống kê trình độ học vấn của các ứng viên theo giới tính

gender	education_level	
Female	Graduate	571
	Masters	244
	Phd	42
Male	Graduate	6095
	Masters	2260
	Phd	217
Other	Graduate	73
	Masters	16

Bước 3: Tính tỉ lệ trình độ học vấn



Bước 4: Tương quan giữa trình độ học vấn và chỉ số phát triển của thành phố nơi mà các ứng viên làm việc

```
gender
Female    0.849370
Male      0.839467
Other     0.873337
```

Như vậy, giới tính khác là các ứng viên làm việc ở các thành phố có chỉ số phát triển cao nhất

32

Đưa ra các câu hỏi có ý nghĩa cần trả lời

Trình độ học vấn và kinh nghiệm làm việc của ứng viên tại các công ty lớn?

- Chúng ta sẽ biết được trình độ học vấn và kinh nghiệm hiện tại của ứng viên đó có xứng đáng để làm việc tại các công ty lớn hay không.
- Công ty lớn sẽ ưu tiên những ứng viên có trình độ học vấn và kinh nghiệm ra sao.
- Công ty lớn có nhận lao động phổ thông nhiều hay không.

Bước 1: Xử lý dữ liệu

- company_size: chuyển các giá trị '10/49' sang '10-49'.
- Ta quy ước: Các công ty có 500 nhân viên trở lên là công ty lớn.

	education_level	experience	company_size
11	Graduate	(1, 5)	5000-9999
12	Graduate	(21, 30)	1000-4999
23	Graduate	(0, 1)	1000-4999
31	Graduate	(16, 20)	5000-9999
34	Graduate	(11, 15)	5000-9999
...	...	...	...
19108	Graduate	(6, 10)	1000-4999
19127	Graduate	(6, 10)	10000+
19132	Graduate	(6, 10)	500-999
19143	Graduate	(21, 30)	10000+
19146	Graduate	(1, 5)	500-999

3209 rows × 3 columns

Bước 2: Nhận xét về trình độ học vấn của lao động ở các công ty lớn

	10000+	1000-4999	500-999	5000-9999
Graduate	0.281	0.178	0.123	0.08
Masters	0.134	0.089	0.048	0.032
Phd	0.013	0.011	0.006	0.005

- Nhân sự chính của các công ty này đa phần là các ứng viên đã có bằng đại học.
- Các công ty quy mô lớn đều sở hữu số lượng Thạc Sĩ, Tiến Sĩ nhất định để đào tạo nguồn nhân lực dồi dào.

Bước 3: Nhận xét về kinh nghiệm làm việc của ứng viên tại các công ty lớn này

	10000+	1000-4999	500-999	5000-9999
(6, 10)	0.1256	0.0701	0.0489	0.0324
(21, 30)	0.1022	0.0754	0.0383	0.0287
(1, 5)	0.0729	0.043	0.0302	0.0171
(11, 15)	0.0729	0.0548	0.0365	0.0231
(16, 20)	0.0511	0.0312	0.0203	0.0165
(0, 1)	0.0041	0.0031	0.0012	0.0003

Bước 3: Nhận xét về kinh nghiệm làm việc của ứng viên tại các công ty lớn này

- Nếu quy ước một ứng viên lành nghề là người có kinh nghiệm từ 5 năm trở lên. Ta sẽ có tỉ lệ phân bố ứng viên lành nghề của các công ty này là:

Tỉ lệ ứng viên lành nghề: 0.828

	10000+	1000-4999	500-999	5000-9999
(6, 10)	0.1256	0.0701	0.0489	0.0324
(21, 30)	0.1022	0.0754	0.0383	0.0287
(11, 15)	0.0729	0.0548	0.0365	0.0231
(16, 20)	0.0511	0.0312	0.0203	0.0165

Tỉ lệ ứng viên không lành nghề: 0.172

	10000+	1000-4999	500-999	5000-9999
(1, 5)	0.0729	0.043	0.0302	0.0171
(0, 1)	0.0041	0.0031	0.0012	0.0003

37

Đưa ra các câu hỏi có ý nghĩa cần trả lời

Sau khi hoàn thành huấn luyện ở các công ty nhỏ, ứng viên có quyết định nhảy việc hay không?

- Sau khi đã trải qua quá trình huấn luyện và làm việc tại các công ty nhỏ, các ứng viên sẽ ưu tiên nhảy việc để tìm bến đỗ tốt hơn hay ở lại và tiếp tục cống hiến
- Các ứng viên kiên trì như thế nào với công ty có quy mô nhỏ như hiện tại.
- Chúng ta sẽ nắm được phần nào suy nghĩ của các ứng viên đó về công ty như: chế độ đãi ngộ nhân sự, thái độ của người sử dụng lao động,...

Bước 1: Xử lý dữ liệu

- Ta quy ước các công ty nhỏ có số lượng lao động không quá 100.

	training_hours	experience	target
1	47	(11, 15)	0
4	8	(21, 30)	0
7	18	(11, 15)	1
8	46	(6, 10)	1
29	68	(16, 20)	1
...	...	...	...
19122	78	(6, 10)	1
19125	31	(0, 1)	0
19135	136	(6, 10)	0
19149	36	(6, 10)	1
19155	44	(21, 30)	0

3636 rows × 3 columns

Bước 2: Nhận xét về tỉ lệ nhảy việc sau thời gian huấn luyện

	training_hours	experience
7	18	(11, 15)
8	46	(6, 10)
29	68	(16, 20)
45	26	(1, 5)
46	87	(1, 5)
...	...	...
19031	87	(11, 15)
19054	31	(1, 5)
19057	56	(21, 30)
19122	78	(6, 10)
19149	36	(6, 10)

781 rows × 2 columns

Bước 2: Nhận xét về tỉ lệ nhảy việc sau thời gian huấn luyện

- Ta quy ước kinh nghiệm làm việc trên 5 năm là lành nghề

Lao động lành nghề

	Less than 100 hours	Between 100 and 200 hours	More than 200 hours
(6, 10)	0.575	0.578	0.5
(11, 15)	0.221	0.156	0.2
(16, 20)	0.117	0.141	0.2
(21, 30)	0.088	0.125	0.1

Lao động không lành nghề

	Less than 100 hours	Between 100 and 200 hours	More than 200 hours
(1, 5)	0.919	0.893	0.8
(0, 1)	0.081	0.107	0.2

Bước 3: Nhận xét về tỉ lệ ở lại công ty sau thời gian huấn luyện

	training_hours	experience
1	47	(11, 15)
4	8	(21, 30)
56	52	(1, 5)
76	65	(1, 5)
79	4	(1, 5)
...	...	...
19089	27	(16, 20)
19099	135	(11, 15)
19125	31	(0, 1)
19135	136	(6, 10)
19155	44	(21, 30)

2855 rows × 2 columns

Bước 3: Nhận xét về tỉ lệ ở lại công ty sau thời gian huấn luyện

Lao động lành nghề

	Less than 100 hours	Between 100 and 200 hours	More than 200 hours
(6, 10)	0.373	0.381	0.427
(11, 15)	0.259	0.276	0.273
(16, 20)	0.245	0.203	0.227
(21, 30)	0.123	0.141	0.073

Lao động không lành nghề

	Less than 100 hours	Between 100 and 200 hours	More than 200 hours
(1, 5)	0.955	0.958	1.0
(0, 1)	0.045	0.042	NaN

43

Tổng kết

44

Khó khăn	• Chưa thành thạo trong việc sử dụng Git, Github • Khai thác các ý nghĩa từ dữ liệu
Kiến thức học được	• Nắm được cách sử dụng các thư viện numpy, pandas, matplotlib, seaborn • Hiểu rõ hơn về Git, Github • Hiểu rõ hơn về quy trình khai thác dữ liệu

Mục tiêu của nhóm:

- Có thể khai thác được ý nghĩa từ dữ liệu nhiều hơn
- Xử lí các giá trị categorical bị thiếu bằng các mô hình học máy thay vì xóa đi

45

Cảm ơn thầy và các bạn đã lắng nghe