DAT565/DIT407 Assignment 8

Sebastian Miles
miless@chalmers.se

Olle Lapidus
ollelap@chalmers.se

2024-10-24

Problem 1

We answer the selected questions from the article *Datasheets for Datasets* by Gebru et al. [**Gebru:2021**]

Motivation

1). The dataset was created to analyze HR functions, and optimize decision making related to employee management. The dataset can be used to analyze and predict attrition, employee performance. Monitoring diversity, compensation analysis and succession planning.

2). The data was created by Fahad Rehman, a student at Abasyn University of Peshawar, Pakistan. He is a Data scientist and Graphic Designer by hobby, according to his github.com biography.

Composition

5). There is only one type of instance, which is the employee record. Meaning the dataset is structured around employees only.

6). There are 15000 employee records in total in the dataset.

8). Each employee record includes the following 10 attributes:

(a) Satisfaction level
(b) last evaluation
(c) number project
(d) average monthly hours
(e) time spent with company
(f) work accidents
(g) if they are still there
(h) promotion within the last 5 years
(i) department
(j) salary

9). We could use "left" as the target label. Given all of the other parameters the model can predict whether or not the employee is going to leave or not.

15). The data contains whether or not an employee has had a work accident which could in some cases be legally confidential. Such data could be considered sensitive, especially if it is linked to an individual.

16). We are not able to see any kind of offensive, insulting, threatening or anxiety inducing data in the dataset.

17). The data does not identify and sub populations, there are not attributes for gender, age or ethnicities.

18). Even though the data does not have a direct identification of the employees, it is possible to identify if the entry stands out and if the employee is known personally, such as works at this company, etc. But by finding the entry, it is likely that the data was known in order to find it anyways.

19). The employee satisfaction and evaluation are personal to each employee, which makes it sensitive to share publicly. Salary and promotions are often private and sensitive, even though it is not under confidentiality laws. There is also work accidents which could be sensitive information

Collection process

26). There is no mention of an ethical review process.

27). The data was collected from third-party sources, not from individual people. The data was scraped from job positing sites with Selenium.

28). There is no mention whether the the individuals were notified, but most likely they were not since it was taken from a public website. It would require significantly more effort to notify every person.

29). It is not specified whether the individuals consented to their data being collected and used. Given that the data was scraped from a job posting site it is likely that that the individuals were not asked for explicit consent by the data collector. However, The person might have accepted data distribution terms when for example signing up on the website.

Uses

40). The dataset is biased because it might not fully represent all departments and demographics. There is also lack of consent and ethical concerns which might become a problem in the future.

41). Using the dataset to make decisions on hiring, firing or promotion could be inappropriate. The dataset could contain biases which would favor certain people and could be discriminatory.

Problem 2

1). The first issue, is the lack of consent from the individuals. Individuals have the right to know how their data is being used and need give permission for its collection and use.

2). The second issue is that the dataset may contain biases from the collection method. This can lead to unsatisfactory decision-making that negatively affects underrepresented groups. Also if the employees knows for example that the satisfaction level correlates to whether or not they would get fired/promote, there is nothing that stops them from lying about it.

3). The last issue is that the data could be interpreted as sensitive information, such as number of work accidents. Perhaps an individual does not wish to share this. Also the satisfaction level is a very personal and private variable and should be kept to themselves.

Problem 3

(a) It does not follow regulations since Article 6 for eu GDPR states that consent is required to process data.

(b) According to article 6c in the eu GDPR, processing is okay if it's necessary for legal obligation. And if cheating is considered illegal then it would be okay to process it.