Counterfactual Explanations for Employment Services

Raphael Mazzine, Sofie Goethals, Dieter Brughmans, and David Martens

Dept. of Engineering Management, University of Antwerp, Belgium
Raphael.MazzineBarbosaDeOliveira@uantwerpen.be

1 Introduction

The employment research field is vast and complex [1, 6], with a large amount of literature dedicated to understanding which factors affect employability [5, 7] and which actions can be taken to enhance employability, either in general or for specific groups [8]. In the same perspective, the use of machine learning represent an opportunity to deal with those challenges [18], specially for large employment institutions which have massive amounts of data on both employers and candidates available that can be applied to various employability enhancement measures. In the literature, there are examples of machine learning models that predict how likely a person will be employed [16, 3, 4], and job recommendation systems [13] that give suggestion to users given her/his characteristics.

A key issue with modern AI models, is their 'black box' nature, meaning that predictions are very difficult, if not impossible, to explain [9]. The field of explainable AI [9, 21] has raised a number of methods that aim to provide such explanations (by ironically adding another layer of complex explanation algorithms). In this paper, we focus on a popular instance-based explanation approach: the counterfactual. A counterfactual explanation provides the minimum set of input features that have to be changed in order to change the predicted class of an observation. For example, a person that was denied for a job position can have as counterfactual explanation things like: *If your skills would include Python and AWS, and if you would also add Dutch as one of the languages you master, then the prediction would change to suitable for this data science position.* They provide explanations for individual predictions [19], and have emerged as a popular and promising tool to explain individual predictions as they are simple, easily understood by people [17] and have key advantages over feature importance methods like SHAP [10] and LIME [15]. Applied to the employability field, a counterfactual could not just help in the task of making black box models understandable, but they could also be used for individual career advice or to find potential biases in automated decision algorithms.

We find eight use cases for the use of counterfactual explanations in an employment context, with different requirements, methodological implications and relevant stakeholders, as illustrated in Figure 1. Using a dataset of over 12 million job listings and more than 3.5 million resumes from a Belgian employment institution, we demonstrate how these explanations can lead to valuable insights for various stakeholders.

2 Methodology

The dataset is obtained from VDAB, which is a public employment service in Flanders (Belgium). This dataset contains detailed information about both the job seekers and vacancies profiles. In this paper we focus on the structured part of the dataset consisting of the skills people can possess. These skills consists of three main categories: studies, competences and languages. The studies include information on all recognized academic degrees in Belgium, ranging from high school degrees to bachelors, masters and PhD's. Competences consist of more specific skills such as mastering certain software applications or coordination a team. Finally, there is detailed information about the languages and their corresponding level.

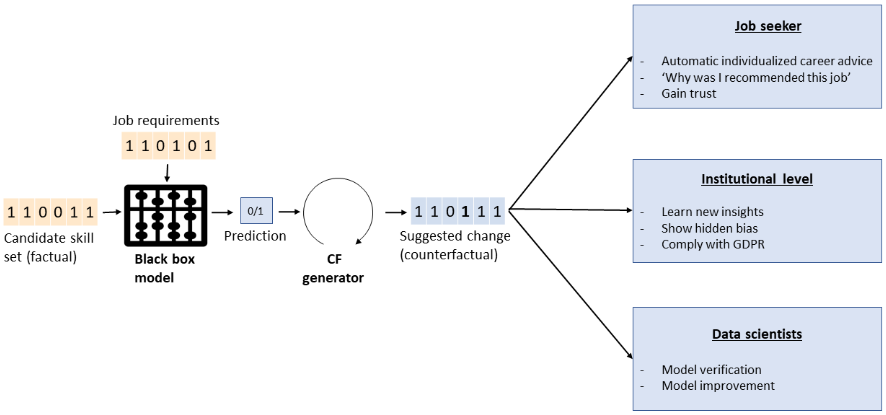

Fig. 1. Schematic overview of counterfactual generation and use cases.

The goal is to build a model that predicts how employable a certain job seeker is for a specific vacancy. As labels are missing, a weighted similarity matching prediction model is built that scores the match between a job posting and a resume, with weights determined by the popularity of a skill. Note that the black-box nature comes from the very large dimensionality of the data, as well as from the possibly non-linear prediction models that are commonly used [11, 14]. We use SEDC [11] as a model-agnostic counterfactual generating algorithm, shown to provide explanations in an efficient manner. Due to its model-agnostic nature, SEDC can be applied to any model which guarantees that it can also be used for any future model that is deployed for the prediction of fits between jobs and resumes.

The generation of counterfactual explanations for a random sample of 1,000 job seeker resumes took on average 1.02 seconds, while the average size of an explanation (the number of changes to modify the prediction from not-employable to employable) was 2.48 features. These numbers are well within the foreseen constraints of practical use. To avoid any confidentiality issues, the provided explanations below are not necessarily built on the obtained dataset.

3 Use cases

3.1 Use cases for the job seeker

Automatic individualized career advice Employment is one of the main issues graduates have to face every year [16], so insights in how one could improve its prospects on a job has large personal, social and economic value. Counterfactual explanations could be used to give personal career advice to individuals searching for a job, where these explanations highlight which skills one would have to learn to get access to the job(s). Consider a person with a degree in Business Administration and a job posting for Business Analyst. The counterfactual explanation shown below, illustrates that the user would be recommended to add the skills Excel and Tableau.

> If you would also have the skills *Excel* and *Tableau*,
> then the prediction would change from not suited for the job to suited for the job.

’Why was I recommended this job?’ Similarly, a job seeker will be presented with some vacancies, by an employment service using automated prediction models, and would like to know why this is the case, surely when the job does not seem well aligned with the profile of the job seeker. An explanation can reveal why this is the case. Unlike the previous use case, in this counterfactual explanation, skills can be removed from the resume (whereas previously we only looked at adding skills in order to change the class). These explanations answer the questions: ’Why was I recommended this job? Which skill do I have to *remove* to no longer receive this recommendations?’ The below explanation reveals that adding skills to your resume that are irrelevant for the job at hand (for example an IT position), can negatively influence your prediction. Such an explanation easily reveals what not to put on your resume (and potentially also, what skills to add).

> If you would not have the skills *carpentry* and *sowing*
> then the prediction would change from not suited for the job to suited for the job.

Gain trust and social acceptance into the model A more general use case for job seekers is obtaining trust in the model and employment service system at large. In general, when machine learning algorithms are used to make (or suggest) any decision to an user, trust is a crucial aspect to consider, specially when dealing with sensitive topics such as employability. People are more likely to accept a machine learning model, if their decisions are understandable [2]. Counterfactual explanations can be used as a mean to gain user trust [23] and, therefore, enhancing the acceptance and impact of complex machine learning systems.

3.2 Use cases for the institutional level

Show hidden bias present in the data Sensitive attributes are often present in the dataset that employments models are trained on, think for example about the well-known case of the automated Amazon recruitment system [12]: the predictive model, trained on historical data, was built to predict whether a candidate was suitable for an engineering position, based on the words of the resume. The model seemed to have a high accuracy, yet it had learned a bias against women by downgrading all-female universities or the occurrence of the word "women's" [12]. Counterfactual explanations could be used to reveal this bias, as shown below.

If you would remove the words '*women's*' and '*all-women university*' from your resume then the prediction would change from not suited for the job to suited for the job.

Learn new insights from the model New insights into the model and the data can be gained by using counterfactual explanations. An example of this is finding population-level reschooling advice. By aggregating the counterfactual explanations for all job seekers (or for a given segment), we can see which are the skills that were recommended the most to job seekers. This could spur reschooling advice, and be valuable for both governmental institutions as for the employment institution itself, as it can offer additional training courses in this area. Figure 2 shows how such analysis could provide both insights for general population and for a specific sector inside population (in this case, people with a Master in Computer Science), where we highlight that the required skills to increase employability can be different depending on the population being investigated.

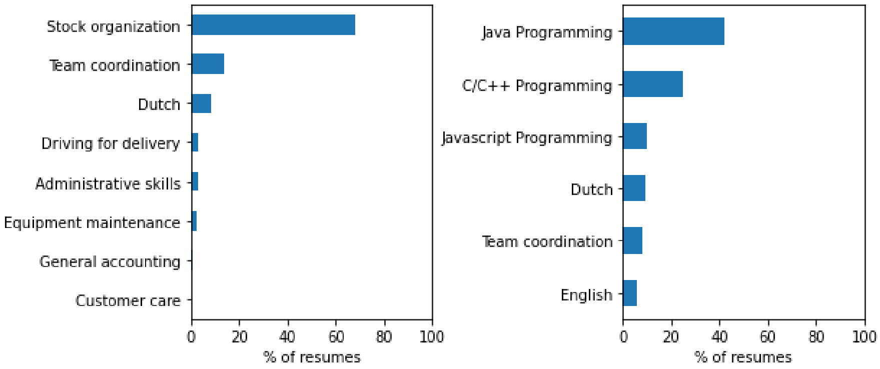

Fig. 2. Conceptual example of most frequently occurring missing skills and languages (features in the counterfactual explanations) of the general population (left chart) and people that have a Master in Computer Science (right chart).

Comply with GDPR regulations If an automated system makes a decision for you that can significantly affect you, which certainly is the case for employment related decisions, a subject has the right to be given an explanation according to the EU General Data Protection Regulation [22]. Counterfactual explanations are a way to comply with this regulation, without opening the black box.

3.3 Use cases for data scientists

Model improvement Due to the high complexity of machine learning models, unexpected classifications (including misclassifications) can occur unknowingly. Counterfactual explanations can reveal the underlying reasons of such cases The below explanation reveals, for example, a resume that is classified as not suited for a given job posting, simply because it is written in French. This likely is due to a bias of few French resumes, which happened to be unsuitable for the listed vacancies. This counterfactual explanation would reveal to the data scientist to either remove or translate these resumes from the dataset, or to add more labeled French resumes.

If you would remove the words '*université*' and '*rue*' from your resume
then the prediction would change from not suited for the job to suited for the job

Verifying the model Finally, counterfactuals are useful in this case by helping to verify why a model made a decision [20], thereby verifying that a reasonable patterns has been learnt. This use cases is similar to the one of trust, but now aimed at the data scientist, with often more specific technical impact. For instance, as part of the MLOps processing pipeline, counterfactuals could be used in several points to verify if any of the changed features are protected or sensitive.

3.4 Conclusion

The eight use cases show the merits of counterfactual explanations for employment services, when using prediction models. These explanations can be used to improve the trust in such automated systems, improve the model performance, dispense career advice, reveal unfair bias, and even provide directions for reschooling efforts. The applied SEDC algorithm is able to provide explanations in a timely manner, quite crucial in this domain where an end user will not want to sit idle in front of a screen, waiting for an explanation. And even though the dimensions of the data are very large (in the thousands), the explanations are quite short and hence understandable for a lay user. In future work, we consider the validation with end users, and further development and adaptation of counterfactual generating algorithms to allow to include constraints on whether to add and/or remove evidence in order to obtain a class change, as different end goals can answer distinct questions.

References

[1] Artess, J., Mellors-Bourne, R., Hooley, T.: Employability: A review of the literature 2012-2016 (2017)

[2] Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)

[3] Casuat, C.D., Festijo, E.D.: Predicting students' employability using machine learning approach. In: 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS). pp. 1–5. IEEE (2019)

[4] Dubey, A., Mani, M.: Using machine learning to predict high school student employability–a case study. In: 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 604–605. IEEE (2019)

[5] Finch, D.J., Hamilton, L.K., Baldwin, R., Zehner, M.: An exploratory study of factors affecting undergraduate employability. Education+ Training (2013)

[6] Guilbert, L., Bernaud, J.L., Gouvernet, B., Rossier, J.: Employability: review and research prospects. International Journal for Educational and Vocational Guidance **16**(1), 69–89 (2016)

[7] Juhdi, N., Pa'Wan, F., Othman, N.A., Moksin, H.: Factors influencing internal and external employability of employees. Business and Economics Journal **11**(1-10) (2010)

[8] Kluzer, S., Rissola, G.: E-inclusion policies and initiatives in support of employability of migrants and ethnic minorities in europe. Information technologies & International Development **5**(2), pp–67 (2009)

[9] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2021)

[10] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in neural information processing systems. pp. 4765–4774 (2017)

[11] Martens, D., Provost, F.: Explaining data-driven document classifications. MIS quarterly **38**(1), 73–100 (2014)

[12] Mujtaba, D.F., Mahapatra, N.R.: Ethical considerations in ai-based recruitment. In: 2019 IEEE International Symposium on Technology and Society (ISTAS). pp. 1–7. IEEE (2019)

[13] Paparrizos, I., Cambazoglu, B.B., Gionis, A.: Machine learned job recommendation. In: Proceedings of the fifth ACM Conference on Recommender Systems. pp. 325–328 (2011)

[14] Ramon, Y., Martens, D., Provost, F., Evgeniou, T.: A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. Advances in Data Analysis and Classification **14**(4), 801–819 (2020)

[15] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

[16] Saouabi, M., Abdellah, E.: Proposition of an employability prediction system using data mining techniques in a big data environment. International Journal of Mathematics & Computer Science **14**(2), 411–424 (2019)

[17] Sokol, K., Flach, P.A.: Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety. In: SafeAI@ AAAI (2019)

[18] Tan, K.L., Rahman, N.A.A., Lim, C.K.: Systematic review in the landscape of data mining and predictive analysis on employability. In: AIP Conference Proceedings. vol. 2016, p. 020142. AIP Publishing LLC (2018)

[19] Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596 (2020)

[20] Vermeire, T., Martens, D.: Explainable image classification with evidence counterfactual. arXiv preprint arXiv:2004.07511 (2020)

[21] Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review. arXiv preprint arXiv:2006.00093 (2020)

[22] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)

[23] Weitz, K., Schiller, D., Schlagowski, R., Huber, T., André, E.: " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. pp. 7–9 (2019)