The Legal and Ethical issues Workshop @LREC2022, pages 38–43
Marseille, 20 June 2022

Transparency and Explainability of a Machine Learning Model in the Context of Human Resource Management

Sebastien Delecraz[1], **Loukman Eltarr**[1], **Olivier Oullier**[2,3,4]
[1]Gojob AI Lab, Aix-en-Provence, France
[2]Inclusive Brains & BRAINS4, Marseille, France
[3]Aix-Marseille University, CNRS, Cognitive Psychology Laboratory (UMR 7290), Marseille, France
[4]Optivio Inc, Boston, Massachusetts, USA
sebastien.delecraz@gojob.com, loukman@gojob.com, olivier@inclusive-brains.com

Abstract

We introduce how the proprietary machine learning algorithms developed by Gojob, an HR Tech company, to match candidates to a job offer are as transparent and explainable as possible to users (i.e., our recruiters) and our clients (e.g. companies looking to fill jobs). We detail how our matching algorithm (which identifies the best candidates for a job offer) controls the fairness of its outcome. We have described the steps we have taken to ensure that the decisions made by our mathematical models not only inform but improve the performance of our recruiters.

Keywords: Fairness, Trust, Human Resources Technologies, Artificial Intelligence, Equal Opportunity Framework

1. Introduction

Human Language Technologies had a significant impact on the business of Human Resource Management (HRM) over the past twenty years. Human Resources Technologies (HR Tech), for instance, have leveraged mathematical models to improve (job) recruitment-related tasks. There are now very efficient models to execute Natural Language Processing (NLP) tasks. These are well suited to process and make sense of the wealth of data (CV, resume, emails, text messages, spoken conversations) that is being exchanged between candidates and employers, including when a recruiter or a recruitment agency acts as an enabler. If one takes the example of data found in resumes, unless guidelines are given to the candidates by the employer or the recruitment agency/platform, most of the time the content to process is not structured. Depending on regulatory constraints (e.g., data protection and privacy laws) in the country in which the recruitment process takes place, as well as the agreement signed by the job candidate prior to sharing data as part of his/her job application, the content of the resume can or cannot belong to the private domain. Regardless of this data being considered private or not, its analyses as part of HRM processes must meet several criteria including (but not limited to):

- an ethical, fair, non-discriminatory and inclusive job selection process;
- transparency and explainability of the mathematical models and non-algorithmic processes employed to assist with decisions;
- compliance to legal and regulatory constraints related to data privacy and protection.

Machine Learning (ML) algorithms have become a key part of decision-making solutions across a great variety of research and business sectors where large amounts of structured or unstructured data need to be processed and make sense of to inform the choices to be made. Today, the way the most efficient ML models (like deep learning or gradient boosting) function is often difficult to monitor. It is also challenging to understand how the algorithm(s) at play make decisions.

If one take the example of gradient boosting algorithms, they are quite opaque, to say the least, in the way they operate. An important issue in the research and business sector leveraging ML is therefore to understand the decision processes and outcomes of the mathematical models at play and which covariates are really acting as discriminators. The challenge of understanding the "algorithmic ghost in the machine" has been picked up by a consortium of multidisciplinary scientists from various country who founded the field of machine behavior: an approach consisting of using rigorous behavioral analytics and metrics to track the behavior of algorithms in order to identify how they make decisions (Rahwan et al., 2019).

In the context of job recruitment, the decisions made by ML models must be controlled, adapted and consistent with the different challenges and objectives of the individuals and/or organizations using them, as well as complying with legal and regulatory constraints. In the HR tech business, the outcomes of ML algorithms must be aligned with the business sector's best practices. This bears a legitimate question of trust and understanding, when compromise between interpretability and performance is too often the name of game. This constitutes a serious issue when, at least in theory, no compromise should be made when it comes to clarity of the data analysis process, ethics, and compliance.

As a temporary recruitment agency leveraging Artificial Intelligence (AI) to optimize its job matching services, Gojob developed a proprietary job matching machine learning solution consisting in a scoring algorithm able to identify the most relevant temporary workers for a request made by one of its clients (i.e., a job offer). Our algorithms are a tool for recruiters to help them staff specific HR needs as fast and as accurately as possible. It is therefore essential for our recruiters to (i) know why a candidate's profile is put forward in (and by) the learning mathematical model, (ii) to understand on which characteristics the recommendation decision is being based, and (iii) to make sure that ML algorithms operate in an ethical, inclusive and therefore non-discriminatory fashion. These are a must-have for the recruiter to trust the ML-powered tools (s)he uses on a daily basis to assist with the decisions to be made to deliver on his/her job. In addition, the recruiter has to be able to justify to the job candidate and to the client (i.e., the possible future employer of the candidate) why a person is deemed fit or not for the position to be filled.

An algorithm should only be considered in light of its performance and results according to a given set of (more or less standard) metrics, but also while taking into account the context in which the data processing it operates (i.e., the decision it makes) happens. This is why, here, in the first section, we introduce some of the safeguards we put in place to ensure that our algorithms, at the core of the daily jobs of our recruiters, do not provide predictions containing ethical and discriminatory biases. In the second section, we show how we used a tool based on the concept of the Shapley values (Shapley, 1953) to reach an acceptable level of accuracy and explainability of the behavior of our Machine Learning models with respect to the different features we use for it to deliver, and keep learning.

2. Ethics and Social Artificial Intelligence

The mission of our company is to provide access to employment to those who are seeking a job, and to offer them the ability to thrive by learning new skills, regardless of their age, gender, origin, education, or level of professional experience. It is also our mission to provide our clients with the best job applicants for the positions they need to fill. Non-discrimination and limited opportunities to learn are major issues blue collar workers face on a daily basis. To date, there is no satisfactory solution available to address this issue, either in Europe or in the US, where Gojob is located. This is what lead to our strategic business decision to have a specific focus on young individuals who are Neither in Employment nor in Education or Training (NEET) in the retail, logistics, and manufacturing industries. There are unfortunately over 2 million people referred to as NEET in France, 10 million in the United States of America (OCDE, 2017). We use our technology to ensure a successful first work experience (or return to work) with our client (i.e., future employer) through training, mentoring, mobility and financial services, while measuring the results after six months based on a set of given criteria: namely the NEET has worked more than sixty days over a period, has signed a contract for a temporary job that lasts more than thirty days, has created their own job or is in training. In 2021, our company has staffed 43% of NEET people as part of temporary job missions, out of approximately fifteen thousand temporary workers (who work on average 5% more hours than other temporary workers).

Given this context, we want to ensure that no particular group is discriminated by our mathematical models and algorithms. Our proprietary database is constituted of applications made by temporary workers in France who are voluntarily and willingly applying for jobs. The atomic item is composed by the set of attributes related to a temporary worker, the set of information related to the job description to which the candidate could (or would) apply for and a label which describes the outcome of the application.

For example, our hypothesis is that applicants who need a residence permit to be allowed to work are more likely to be negatively affected by the model because of a possible bias in our database. The same thing goes for other sensitive attributes (age, gender, nationality, etc.). Therefore, we use dummy variables to categorize, detailed in previous works (Delecraz et al., 2022), what we assume would be a group of candidates that would be "favored" by the algorithm, as opposed to the group that would be "discriminated" by:

- **gender**: male or female;
- **nationality**: French nationality or not;
- **place of birth**: born in France or not;
- **education**: has declared an education level or not;
- **residence permit (RP) requirement**: can work without a residence permit or need to have one;
- **age**: four age groups (18–25, 25–35, 35–45, 45-55) that we consider independently of each other (given a group, we compare those who belong to it against the rest of the population). We stop our ages groups at 55 years old because the number of candidates in our database older than 55 years is way too low (mostly because this age group is generally not seeking temporary jobs) to conduct a qualitative analysis.

We conducted an analysis across these sensitive attributes to assess the fairness of the outcomes provided by our ML model (based on regularizing gradient boosting using XGBoost, an optimized distributed gradient boosting library) using the FairLearn toolkit (Bird et al., 2020), an open-source project which provides

which proposes many methods of fairness analysis for machine learning models. We examined how the model performs based on the Equal Opportunity fairness definition; a metric considered in the specific scientific literature (Hardt et al., 2016) to be the more relevant one to address this question. If $\hat{Y}$ is a binary predictor of the outcome of a worker application and Y the associated ground truth, we consider the class 1 as the preferred outcome in the classification task (the worker was recruited). Given a sensitive attribute A indicating the belonging to a group considered as discriminated and $\bar{A}$ the belonging to the favored group, $\hat{Y}$ is considered equal opportunity with respect to sensitive attribute A if:

$$\mathbf{P}\left(\hat{Y} = 1 \mid Y = 1, A\right) = \mathbf{P}\left(\hat{Y} = 1 \mid Y = 1, \bar{A}\right) \tag{1}$$

Before implementing a fair algorithm, we analyzed the data to observe possible biases towards and/or underrepresentation of some categories. We observed that the distribution of the label is not the same across sensitive attributes. All the work related to this subject is detailed in a previous article (Delecraz et al., 2022). Our analysis shows that our model never exceeds the 5% True Positive Rate Parity (which is the absolute value of the difference between the two probabilities in equation 1). Of course there is no threshold that defines if a model is fair or not. In the literature, depending on the application, we find references to thresholds ranging from 5% to 20%. In our case, we take these first scores as a starting point and of course aim at a score of 0%.

3. Explainability of the Outcomes of Machine Learning Model

Machine learning models are designed and used to optimize a metric or a cost function. In the case of our ML-based solution to assist in job recruitment tasks, knowing that our model considers a candidate as relevant to a given offer is not enough. We need to give the recruiter a minimum amount of insights and information when (s)he reviews the profile of a job candidate, in order to understand which variables were particularly discriminating, either locally or overall. Our intent is to actually understand the rules the algorithm has generated, not just how the algorithm functions. Understanding a model consists in analyzing how it works as a whole, in a given context, that is to say through the input data, the algorithm itself, the output predictions, the weights the model gives to different features, the distributions of different variables and the effect the model gives each one.

It is therefore important to know the "why" of a prediction and to identify the instances where the model is flawed. In particular, the model can make (rightly or wrongly) unexpected decisions, and it is therefore essential to understand what has influenced the prediction in one direction and what could have influenced it to go the other way. A better understanding of the model can sometimes lead to a better understanding of the problem (or question to answer) and to the discovery of new subtleties. Molnar (2020) give a good explanation and broader vision of the explainability issues at play.

Machine Learning techniques are destined to become more and more widespread and to intervene very regularly in decision-making processes in professional and personal settings. Today, the most efficient models are not easy to interpret and there is a lack of visibility on how their decision processes operate. For example, gradient boosting algorithms are quite opaque. An important issue is to understand the decisions output of our model and which covariates are really discriminating. In a context focused on recruitment, the decisions made by the model must be controlled, adapted and consistent with the different challenges. It should be noted that the decisions must be aligned with the business knowledge. This is a true question of trust and understanding, and the compromise between interpretability and performance is not always obvious. We used the SHAP library toolkit (Lundberg and Lee, 2017; Lundberg et al., 2018; Lundberg et al., 2020), based on the concept of Shapley values (Shapley, 1953). This tool allows one to have a local and global vision of the decisions of the model in an agnostic way. The SHAP value measure the participation of a feature to the prediction. For each prediction of our model, we compute the SHAP value for each of its feature.

3.1. Global and Local View

We have adopted two different observation positions to try to explain the decisions of our model. A close up view explanation can provide a clear view. In particular, a global view can give the impression of complex dependencies on a given covariance, whereas a local view can show simpler and clearer interactions. This allows one to see what might change in the output if the input changes slightly. Yet, a helicopter view allows one to access aggregated information as well as to get an idea of how the model works on a group rather than an individual. It is possible to group instances according to the granularity we want to consider.

3.1.1. Global view

The global view allows us to quickly understand which features matter when the model makes a decision. The set of features we use in our model is designed to capture the different characteristics that allow us to evaluate the suitability of a temp for a job offer. The SHAP library provides a tool to examine global model behavior. We report the SHAP importance compute for each feature in Table 1. A SHAP value is computed for all feature for each prediction. Given a feature, we compute its importance by doing the average of the absolute values of the SHAP values overall the predictions. These values allow us to identify the features that, overall, have the most impact. Features meaning and name

have been deliberately hidden for reasons of confidentiality.

Features	SHAP importance
Feature 13	0.85489
Feature 2	0.78975
Feature 7	0.48765
Feature 3	0.25791
Feature 12	0.24747
Feature 8	0.16531
Feature 0	0.16264
Feature 1	0.15541
Feature 5	0.05532
Feature 9	0.03777
Feature 11	0.02633
Feature 4	0.02554
Feature 10	0.01142
Feature 6	0.00768

Table 1: SHAP importance for each feature used by the model.

The explainability that results from these first figures in this global view is more useful to the teams that design the model than to the end users, namely our recruiters. However, further analysis allows us to learn more about the different features, especially those that have a low importance in decision-making. For example, we can deduce that features with a very low SHAP value do not capture well enough what can be decisive for a recruitment and that they should either be improved or removed from the model. However, we can also observe one of the weaknesses of the SHAP tool, namely the existence of correlation between features. For example here, we have calculated that on our corpus of data, Feature 10 and Feature 13 have a correlation score of 0.6. The model (remember that it is an XGboost) could therefore have identified this correlation and decided not to give importance to Feature 10, as Feature 13 would allow it to obtain more or less the same decisions.

In Figure 1 we can identify the SHAP values for each variable. On the x-axis we can see their impact (on the right if the impact on the prediction is positive, on the left if it is negative). On the y-axis are printed the different variables and ordered by the total magnitude of their SHAP values. The color code indicates the value of the variable (the closer the color is to red, the higher the value, conversely blue symbolizes low values). Note that when several points are aligned horizontally but scattered vertically, they represent points that have been impacted similarly. Whereas points that are horizontally distant but have the same color represent instances that have been impacted differently for similar values of the variable concerned. This last case means in particular that there was an interaction with other variables.

By reading the feature importance in Figure 1, we can first notice the monotonic effects that the model has

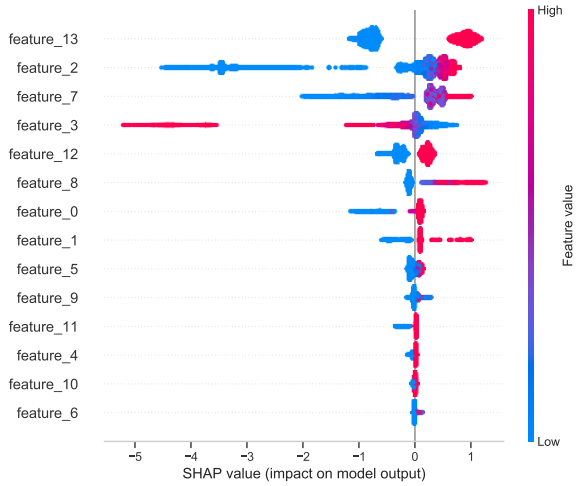


Figure 1: Feature importance with SHAP values. For each dot, vertical position depict the feature, horizontal position indicates whether the effect of that value caused a higher or lower prediction and the color describes the value taken in that dot.

learned for every feature. Feature 13 has the biggest weight here and its effect is quite unequivocal. The fact that the points are located in two small clusters show little interaction with other features and a strong homogeneous effect. The Feature 2 shows a more uneven effect. Its high values have a similar impact on the model, but the low values are more spread out. Some latter have a more or less neutral impact, but some others have a quite strong negative impact and take a wide range. The Feature 2 strong importance is likely to come from these instances rather than a really global effect. Feature 3 has a similar behavior to the Feature 2. However, it differentiates itself because of its negative monotony and also because some really high values endure a negative impact that is really far from the one on the other points. In contrast, Feature 4, 10 and 6 are meaningless for the model. The model ignores those features, and we can say that they have little effect on the model decisions.

We could go further in the plot reading, and it really is a wealthy source of information. One only has to get a good understanding of the global effects. It is possible to discern whether a feature effect is global or focused on some instances only. Moreover, a very important aspect is to challenge these effects and make sure they are aligned with the business logic. Depending on the case a model should not base its decision on one feature only but rather on interactions and non-linear effects. This action permits us to find undesirable effects and debug the model.

3.1.2. Local view

We also analyze the insights on local prediction and visualize the effect of the different variables. In the Appendix A, we provide a couple of examples we randomly chose in the data. In the Figures 2a we will zoom

in on some predictions. The red color indicates positive impact and the blue color indicates a negative impact. Each time we can see the value of the concerned feature. We also see the output value (negative value means the model gave the negative class to the instance x).

If we compare Figure 2a with Figure 2b we can see that the outputs are appreciably similar but the reasons are totally different. Both instances have been given the negative class by the model. In the Figure 2a the feature that contributes the most is Feature 7 followed by Feature 13 and 8. Here the negative impact was carried by three features only. The other variables had a mild positive impact. To extend the reflection conducted in the global view part, Feature 2 (which had an increasing effect on the model) has a positive impact with a value of 0.229. This value can seem low, but the model estimated that it had a positive impact. Therefore, a worker corresponding to this instance should increase his value for the Feature 2 and also for Feature 7 if he wants to be classified as positive the next time. In the Figure 2b, the most important feature was Feature 13 as it was in the previous example. However, other features have been reversed as Feature 12, 7 and 2. One can see that according to these two examples, turning Feature 7 from 0.025 to 0.245 has a stronger impact than turning Feature 2 from 0.229 to 0.005.

In Figure 2c, we can see that the profile was supported by the model because of the high values for both of features 13 and 7. In the three examples Feature 8 was low and brought a really mild negative impact. To go further we could explore the distribution of this variable and in what scenarios it has a positive impact eventually.

With this type of representation, we can quickly explain to the recruiters which features have the most impact on the model's decision. In case of rejection, our recruiters have the possibility to send feedback to the candidate to explain the reason, or the actions that the candidate can do to quickly increase his chances to be qualified for the job offer. In case of acceptance, the recruiter has explanations about the decision-making which improves his confidence in the model. He can also provide a detailed explanation to the client on the relevance of the candidate to the job offer.

4. Conclusion

The study in this article shows the control we maintain over our Machine Learning algorithms as a social impact company. While no final hiring decision is made by a machine alone, the choices the matching algorithm makes must be fair and explainable to our recruiters for them to make the final call. This is why we have built into our AI-based automation process algorithmic safeguards that signal possible biases (theory) and measured biases (outcomes) as well as ways to visualize and understand what sources of information the model's decisions are based on. We strongly believe that safeguards algorithms to minimize biases and discrimination should become the norm when artificial intelligence is used in job recruitment processes.

5. Bibliographical References

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May.

Delecraz, S., Eltarr, L., Becuwe, M., Bouxin, H., Boutin, N., and Oullier, O. (2022). Making recruitment more inclusive: Unfairness monitoring for a job matching machine learning algorithm. In *International Workshop on Equitable Data and Technology (Fairware'22)*, pages 364–372, Pittsburgh, Pennsylvania, USA, 5.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

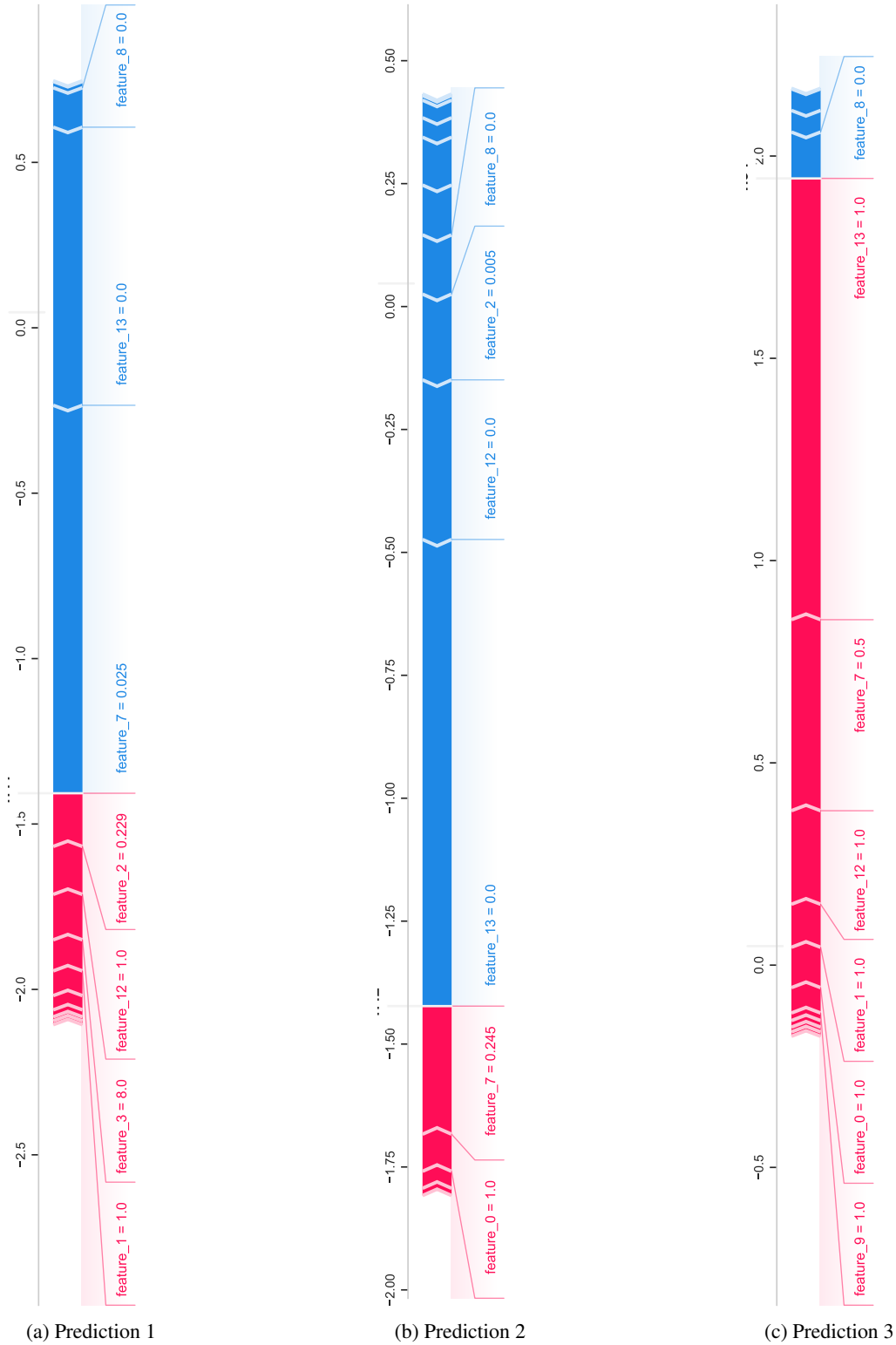
Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

OCDE. (2017). Youth not in employment, education or training (neet).

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, 568(7753):477–486.

Shapley, L. S., (1953). *A Value for n-Person Games*, volume 2, chapter 17, pages 307–318. Princeton University Press.

A. Example of SHAP value decomposition for predictions



(a) Prediction 1 (b) Prediction 2 (c) Prediction 3

Figure 2: Decomposition of two negative and one positive prediction showing each feature impact with SHAP values.