Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI for Good

Toward Job Recommendation for All

Guillaume Bied[1,2] , **Solal Nathan**[1] , **Elia Perennes**[2,3] , **Morgane Hoffmann**[2,3] , **Philippe Caillou**[1] , **Bruno Crépon**[2] , **Christophe Gaillac**[4] and **Michèle Sebag**[1]
[1]Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), Orsay, France
[2]Centre de Recherche en Economie et Statistique (CREST), Palaiseau, France
[3]Pôle emploi, Paris, France
[4]Oxford University, Oxford, United Kingdom
bied@lri.fr

Abstract

This paper presents a job recommendation algorithm designed and validated in the context of the French Public Employment Service. The challenges, owing to the confidential data policy, are related with the extreme sparsity of the interaction matrix and the mandatory scalability of the algorithm, aimed to deliver recommendations to millions of job seekers in quasi real-time, considering hundreds of thousands of job ads. The experimental validation of the approach shows similar or better performances than the state of the art in terms of recall, with a gain in inference time of 2 orders of magnitude. The study includes some fairness analysis of the recommendation algorithm. The gender-related gap is shown to be statistically similar in the true data and in the counter-factual data built from the recommendations.

1 Introduction

Machine learning is increasingly used in the domain of human resources, tackling e.g. the recommendation of career paths [Geyik *et al.*, 2018; Ramanath *et al.*, 2018; Shalaby *et al.*, 2017] or the identification of "churners" [Sisodia *et al.*, 2017; Yadav *et al.*, 2018]. This paper focuses on e-recruitment, i.e. the design and exploitation of recommender systems selecting job ads best suited to job seekers.

As noted by [Fernandez and Gallardo-Gallardo, 2020; Mashayekhi *et al.*, 2022], e-recruitment (e-R) faces specific challenges compared to the recommendation of goods on e-commerce platforms. Firstly, due to the sensitivity of the e-R data, the domain lacks a comprehensive open benchmark supporting the comparative assessment of the algorithms, playing a similar role as ImageNet for computer vision [Russakovsky *et al.*, 2014]. The RecSys 2016-17 challenge data [Abel *et al.*, 2016; Abel *et al.*, 2017], to our best knowledge the most representative dataset for e-R, considers a simplified hierarchy of the job sectors and career levels and is deprived of geographical information due to privacy concerns. Secondly, e-R involves rival goods (a single job can be attributed to a single job seeker), with an extremely sparse interaction matrix, and recommendation mostly considers new job seekers and recent job ads (cold-start recommendation). Both features increase the complexity of the e-R problem, e.g. hindering the extraction of latent representations from the interaction matrix, or requiring specific methods to exploit such latent representations in cold-start mode [Volkovs *et al.*, 2017b]. Additionally, the description of the data is heterogeneous (text; list of skills; past employment for job seekers). Thirdly, due to the impact of an e-R system on people's lives, specific care is required to enforce the fairness of the recommendations and account for the potential biases in the data [Islam *et al.*, 2021].

This paper presents an e-R system called *MUlti-head Sparse E-recruitment* (MUSE) learned from proprietary data of the French public employment service *Pôle emploi*, featuring two other specifics compared to the e-R state of art (section 2). Firstly, *Pôle emploi* aims to serve each and everyone, and in practice mostly serves low-qualified job seekers (paid at circa the minimum legal wage); the targeted audience is associated with a less informative description and more sparse interactions compared to social network-centered approaches [Ramanath *et al.*, 2018] (more in section 2). Secondly, the sought e-R system must be scalable and able to serve millions of job seekers facing some hundred thousands job ads.

The contributions of the MUSE approach (section 3) are twofold. On one hand, MUSE favorably compares with the state-of-art [Volkovs *et al.*, 2017a] w.r.t. the standard recall performance indicator, with a gain of 2 orders of magnitude in inference time (section 5). A first online testing of the approach on 20,000 job seekers has confirmed the acceptability of the recommendations compared to a preference-based system inspired from the proprietary *Pôle emploi* system (section 6.1). MUSE is learned as a 2-tier neuronal architecture. The first tier, MUSE.0 operates a fast selection of the top 1,000 job ads for each job seeker, where specific embeddings are learned to model geographical and skills aspects. On the top of the MUSE.0 filter, the second tier exploits a refined and more expensive representation of job seekers and job ads. It learns an agnostic model of the matches (i-th job seeker is hired on j-th job ad) and of the applications (i-th job seeker applies on j-th job ad), and an end-to-end model combining both hiring and applications.
On the other hand, the fairness of the recommendation model is thoroughly investigated (section 6.2). The proposed methodology compares the gender-related biases in the actual hirings and in the counter-factual framework of the recommended hirings, showing *no statistically significant increase*

of these biases. The paper concludes with some perspectives.

2 Related Work

e-R has gradually emerged as a major domain of "AI for good" [Xiao *et al.*, 2016; Volkovs *et al.*, 2017a]. Referring to [Freire and de Castro, 2021; Fernandez and Gallardo-Gallardo, 2020; Mashayekhi *et al.*, 2022] for a comprehensive survey of the domain, this section discusses the approaches most relevant to the context of an e-R system for a public employment service.

Quite a few works are centered on a social network dedicated to employment and careers, such as LinkedIn [Li *et al.*, 2016; Geyik *et al.*, 2018; Kenthapadi *et al.*, 2017; Ramanath *et al.*, 2018; Borisyuk *et al.*, 2016; Zhang *et al.*, 2016; Ozcaglar *et al.*, 2019] or CareerBuilder [Zhao *et al.*, 2021]. Compared with the wide audience of *Pôle emploi*, the higher homogeneity of the social network audience induces a job seeker distribution that is both more compact and more informative (e.g. with more diversified skills).

As said, the main public benchmark[1] contributed by Xing and used for the RecSys 2017 challenge [Abel *et al.*, 2017] is transformed to ensure the data privacy, using manual feature construction, pre-processing textual data and removing geographical information. Among the prominent approaches developed for or validated on the RecSys dataset are [Volkovs *et al.*, 2017a] and [Volkovs *et al.*, 2017b]. In [Volkovs *et al.*, 2017a], an XGBoost [Chen and Guestrin, 2016] algorithm is optimized to the challenge scoring function. In [Volkovs *et al.*, 2017b], the rich interaction matrix is exploited to extract a latent representation of the users, and align the embedded representation of the brand new users to handle the users with no previous interactions (cold start recommendation). In [Yagci and Gurgen, 2017] a ranker ensemble is used in the same context.

The quality and sparsity of the interaction matrix also is a key difference among the RecSys dataset and standard *Pôle emploi* datasets. In the RecSys case, a variety of interactions among job seekers and job ads (e.g., view, click, bookmark, reply, recruited) are reported, and the scoring function associates a weight with each interaction correctly predicted. In the *Pôle emploi* case, the only available interactions are "hire" and "apply". The "hire" interaction matrix is a permutation[2] for the job seekers who found a hire, and 0 otherwise; and the the application matrix is almost as sparse with an average of 1.06 application per job seeker. As said, this sparsity adversely affects the extraction of a latent representation using matrix decomposition [Volkovs *et al.*, 2017b]. On the other hand, the geographical information, missing in the RecSys dataset, plays a key role for the recommendation of low-qualified jobs.

The multi-faceted nature of the recommendation, involving the adequacy of the job seeker profile and job ad w.r.t. *e.g.* skills; nature of contract, hours, salary; geographical distance, is accounted for in the early ELISE approach [WCC, 2023] and the *Pôle emploi* system relies on the weighted aggregation of these facets. Some approaches like [Biancofiore *et al.*, 2021] have tried to leverage Knowledge Graphs to use this rich information.

The use of a filter to narrow down the search is presented in [Borisyuk *et al.*, 2016]. Generalized linear mixed models [Zhang *et al.*, 2016] are used to combine user and item features [Ozcaglar *et al.*, 2019]. In [Zhao *et al.*, 2021], a filter is built using a weighted sum of textual and geographical adequacy of the job seeker and job ad, exploiting their rich textual description.

Given the impact of recommender systems on people's life, the fairness of recommendations is increasingly studied [Ekstrand *et al.*, 2022; Wang *et al.*, 2023; Li *et al.*, 2022], considering the fairness w.r.t. users and items (e.g. providing all items a fair exposure to users' attention). In e-R, the main issue is to provide subgroups with fair recommendations, e.g. similar performance [Wang *et al.*, 2023]. The trade-off between the performances (recall) and the fairness of a recommendation policy has been investigated, arguing that recommendations should not depend on users' features such as gender. The magnitude of the gender-related impact is measured in [Li *et al.*, 2021]. In [Rus *et al.*, 2022; Islam *et al.*, 2021; Li *et al.*, 2021], fair recommendations are obtained using adversarial approaches to enforce the neutrality of the representation w.r.t. sensitive attributes.

3 Overview of MUSE

This section presents the 2-tier MUSE architecture and discusses its multi-head structure. The first tier MUSE.0 (Fig. 1) aims to enforce the scalability of the recommendation functionality. Specifically, MUSE.0 uses the elementary descriptions of the job seeker $\mathbf{x}$ and the job ad $\mathbf{y}$, and computes a fast score MUSE.0$(\mathbf{x}, \mathbf{y})$. This score is exploited to rank and filter all but the top 1,000 job ads, narrowing the search for the second tier MUSE.1 and enabling the use of a more complex representation.

3.1 MUSE.0

MUSE.0 models the main three facets relevant to job recommendation, respectively concerned with competences and skills, geographical, and general aspects. The faceted match of job seeker $\mathbf{x}$ and job ad $\mathbf{y}$ is sought as:

$$s(\mathbf{x}, \mathbf{y}) = \langle \phi_0(\mathbf{x}), \psi_0(\mathbf{y}) \rangle$$

where embeddings ϕ_0 and ψ_0 are trained using a triplet loss [Weinberger and Saul, 2009]. Noting $(\mathbf{x}, \mathbf{y}, \mathbf{y}')$ a triplet made of job seeker $\mathbf{x}$, their match $\mathbf{y}$ and another job ad $\mathbf{y}' \neq \mathbf{y}$,[3] the loss is defined as:

$$\mathcal{L}(\phi_0, \psi_0) = \sum_{(x,y,y')} [\langle \phi_0(\mathbf{x}), (\psi_0(\mathbf{y}) - \psi_0(\mathbf{y}')) \rangle + \eta]_+ \quad (1)$$

[1] The CareerBuilder public dataset used for 2012 Job Recommendation Challenge only involves applications (no clicks, no hires). Compared to RecSys, it is smaller and involves lesser interactions though richer textual information.

[2] Except for the tiny fraction of users with several matches, e.g. hired on interim jobs.

[3] $\mathbf{y}'$ is uniformly sampled among the job ads *available during the match week*. More sophisticated negative sampling strategies have been considered with no improvement.

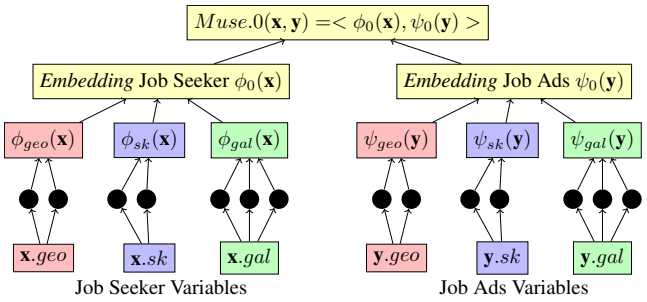

Figure 1: MUSE.0 architecture: three embeddings are defined to model geographical, skills and general aspects of job seekers (left) and job ads (right), and compute the hiring score.

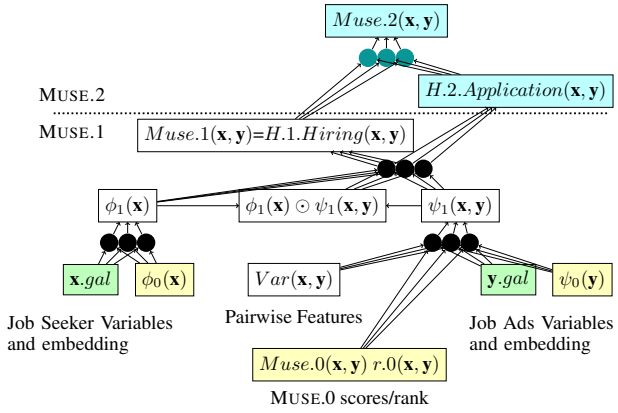

Figure 2: MUSE.1 (below dashed line) and MUSE.2 architectures. MUSE.2 includes a second-head to model the applications, and a top head, exploiting both the standalone hiring and the application scores to predict the overall hiring score.

with $[x]_+ = \max(x, 0)$ and $\eta > 0$ a margin factor.

ϕ_0 and ψ_0 are defined by concatenating three embeddings, respectively reflecting the skills, geography and remaining other information. The use of these three facets, each relevant to matching on the labor market, is empirically justified by ablation studies in section 5.

The skill matching module (ϕ_{sk}, ψ_{sk}). The job sectors are structured along a tree-structure ontology including 14 sectors ("agriculture", "healthcare"), composed of 110 intermediate sectors ("woodcutting and pruning") and 531 detailed types of job. Each type of job is associated with a list of skills using expert knowledge; the recruiter can specify additional skills, required to occupy the job. Job seekers likewise describe their skills and possess by default the skills associated with their sought jobs. The catalog of standardized skills includes circa 12,300 terms (e.g., "welding techniques", "tax system knowledge"). The skill description of both job seeker and job ad is a 12,300 binary vector. Embeddings ϕ_{sk} and ψ_{sk} are learned using a triplet loss.

The geographical matching module (ϕ_{geo}, ψ_{geo}). This module is based on a tiled representation of the locations, taking inspiration from kernel density estimation and matrix factorization [Lian *et al.*, 2014]. Formally, given a reference grid paving the national or regional territory with points z_i, the geographical representation of a job seeker (resp. job ad) situated at $x.geo = z \in \mathbb{R}^2$ and supplied as input of the geographical module is set to:

$$\left(\exp^{-\tau \cdot d(z,z_i)^2}\right)_i$$

with $\tau > 0$ controlling the granularity of the representation (the number of non-zero coordinates) and d the geodesic distance.

Embeddings ϕ_{geo} and ψ_{geo} are learned on the top of this tiled representation using a triplet loss.[4] This module reflects the fact that the impact of the distance of a job seeker to a job depends on other factors (public transportation; traffic jams) than the distance in km: it is not invariant by translation.

The general matching module (ϕ_{gal}, ψ_{gal}). The general module model takes as input a 500-dimension vector with all information related to job seekers (age, required salary and type of contract, textual description) and job ads (offered salary and textual description of the job and of the company). It includes a 50-dimensional SVD representation of the skills, the location (latitude and longitude) and a 100- (respectively, 200-) dimensional SVD representation of the textual description of job seeker (resp. job ad). Embeddings ϕ_{gal} and ψ_{gal} are likewise learned using a triplet loss.

The training schedule. Each module standalone is trained in a first phase; all modules are jointly trained and fine-tuned in a second phase using stochastic gradient descent with Adam [Kingma and Ba, 2014]. Overall, MUSE.0 yields a scalar matching score:

$$\text{MUSE.0}(\mathbf{x},\mathbf{y}) = \sum_{m \in \{sk,geo,Gal\}} \langle \phi_m(\mathbf{x}), \psi_m(\mathbf{y}) \rangle$$

3.2 MUSE.1

As said, the MUSE.0 score is used to filter the job ads considered for each job seeker. The recall@1,000 of MUSE.0 is above 80%, making it possible to only consider the top 1,000 job ads for each job seeker, with a limited loss in recall. MUSE.1, refining the ordering of the top 1,000 job ads, uses more complex features $Var(\mathbf{x},\mathbf{y})$ depending on both job seeker $\mathbf{x}$ and job ad $\mathbf{y}$,[5], which would not be possible for scalability reasons if all available job ads should be considered.

MUSE.1 (Fig. 2) takes as input the description of $\mathbf{x}$ and $\mathbf{y}$ (same as input of the General module), their elementwise product, the crossed features $Var(\mathbf{x},\mathbf{y})$ and the information provided by MUSE.0, i.e. the latent description $\phi_0(\mathbf{x})$ and $\psi_0(\mathbf{y})$, the score MUSE.0$(\mathbf{x},\mathbf{y})$ and the rank of $\mathbf{y}$ by decreasing order of MUSE.0$(\mathbf{x},\mathbf{y})$. Overall, the recommendation score learned by MUSE.1 reads:

$$\text{MUSE.1}(\mathbf{x},\mathbf{y}) = \text{MLP}(\phi_1(\mathbf{x}), \psi_1(\mathbf{x},\mathbf{y}), \phi_1(\mathbf{x}) \odot \psi_1(\mathbf{x},\mathbf{y}))$$

[4] The only difference w.r.t to Eq. 1 lies in the negative sampling, as job ad $\mathbf{y}'$ is uniformly selected among the job ads contemporary of $\mathbf{y}$ *and situated farther away* from job seeker $\mathbf{x}$.

[5] Vector $Var(\mathbf{x},\mathbf{y})$ measures the adequacy of an $(\mathbf{x},\mathbf{y})$ pair *re* the distance, skills, occupation, education, experience, contract type, spoken languages, driving licenses and wages.

where MLP denotes a multi-layer perceptron, and ϕ, ψ are job seeker & job ad embeddings. MUSE.1 is trained by minimizing a cross-entropy loss:

$$\mathcal{L} = \sum_{(x,y,y')} \log(\text{MUSE.1}(\mathbf{x}, \mathbf{y})) + \log(1 - \text{MUSE.1}(\mathbf{x}, \mathbf{y}')) \quad (2)$$

3.3 MUSE.2

As said, a critical difficulty of e-R in the *Pôle emploi* framework is the extreme sparsity of the interaction matrix in the dataset (a single hire being reported for the hired job seekers, and 0 for the others). On the other hand, the dataset records some of job seekers' applications, when they are submitted through the *Pôle emploi* online platform, or mediated by caseworkers. [6]

Accordingly, a multi-head MUSE.2 architecture is considered to enable information sharing between the hiring and the application interaction matrices (Fig. 2, Top). A first head aims to predict the hirings; a second head aims to predict the applications; a third head, aimed to predict the hirings, is learned on the top of both first and second heads, likewise using a cross-entropy loss (Eq. 2).

4 Experimental Setting

A preliminary online testing on 20,000 real users, conducted to assess the acceptability of the recommendations, is described in section 6.1. This section describes the experimental setting used to compare MUSE to the state of the art.

4.1 Goals of Experiments

Our primary goal is to comparatively assess the performance of MUSE in terms of both performance and inference time. The single head (MUSE.1) and the multi-head (MUSE.2) architectures are compared and the impact of the different modules is assessed using ablation studies. Another goal is to inspect and measure the biases due to the recommendation algorithm, and compare these biases with those present in the data (section 6.2).

4.2 Benchmarks

RecSys. As said, the public dataset most relevant to e-R is the dataset released for the ACM Recsys 2017 challenge[7], provided by the social network Xing. It involves 1.5M job seekers, 1.3M jobs and 30M interactions, recorded from Nov. 2016 to Jan. 2017 in Germany, Austria and Switzerland. After thorough anonymization, removal of geographical information[8] and pre-processing of textual data, job seekers and job ads are represented as binary or categorical vectors of dimension respectively 831 and 2,738. The interaction matrix reports 6 levels of interaction, 4 of which (*click, bookmark, reply, recruited*) are interpreted as "hiring". The fifth level (*impressions*) is interpreted as "applying" and used for the MUSE.2 training. The same training/test split and procedures as in [Volkovs *et al.*, 2017b] are used, including: i) a warm start scenario (426K interaction pairs), where users and items involved in the test set are also present in the training set; ii) a user cold-start scenario (159K pairs) where 42,153 test users have no interactions in the training set.

Pôle Emploi This proprietary dataset involves 1.2M job seekers, 2.2M job ads, with an overall number of matches of 242k and 1.29M applications, recorded from Jan. 2019 to Sept. 2022 in a French region. The hiring interaction matrix is embarrassingly sparse: it includes a single interaction for 96.3% of the hired job seekers[9] and 0 for the non-matched job seekers and job ads, preventing the extraction of a latent representation of the data. The application interaction matrix is also significantly sparser than in the RecSys dataset: circa 80% of the job seekers have no application, and 95% have less than 4 applications. The training set includes 85% uniformly selected weeks from Jan. 2019 to Sept. 2022. The test set, including the remaining weeks, involves circa 400k job seekers, 70k job ads and 1.4k matches per week.

4.3 Baselines

The first baseline is a home-made version of the XGBoost winner of the RecSys challenge [Volkovs *et al.*, 2017a] (that is not publicly available, and tailored to optimize the challenge scoring function). On the *Pôle emploi* data, XGBoost is provided with the description input of the $general$ MUSE.0 module ($x.gal$ and $y.gal$) plus the cross-features $Var(\mathbf{x}, \mathbf{y})$ also used by MUSE.1 and MUSE.2 for a fair comparison. The second baseline is DropoutNet [Volkovs *et al.*, 2017b], that exploits both the job seeker and job ad description and their latent description extracted from the interaction matrix.[10] Other algorithms, e.g. [Zhao *et al.*, 2021], that heavily rely on textual and geographical information, do not apply on the considered datasets.

5 Experimental Validation

The reported computational times are obtained on Intel(R) Xeon(R) Silver 4214Y CPU @ 2.20GHz, with 187 GB RAM and a Tesla T4 GPU. Experiments on the *Pôle emploi* dataset are conducted on a secure platform. More detail about the experiments is provided in Supplementary Material.

The results report the recall indicator and the computational time. Significantly best results (with 95% confidence with respect to the second best result) are legended "*" in all tables.

5.1 The RecSys Dataset

MUSE configuration is adapted to fit the specifics of the RecSys dataset (more in supplementary material).

[6] Domain knowledge is leveraged to select the applications on the initiative of the job seeker, or those proposed by a caseworker and approved by the job seeker, expectedly better reflecting their preferences.

[7] http://www.recsyschallenge.com/2017/

[8] Only the country or German Lander are available.

[9] 3.7% of the job seekers are hired on several interim jobs, in the period.

[10] We thank the authors for making the DropoutNet code public, together with the data and the latent description of the RecSys job seekers and job ads.

Recall@100	DN	MUSE.0	MUSE.1	MUSE.2
Warm-start	**41.2***	13.0	24.9	24.9
Cold-start	23.1	12.3	23.3	**24.0***
Training	>10h	2.7h	1.25h	8.3h
Rec. per j.s.	0.001"	0.002"	0.013"	0.016"

Table 1: Comparative results of MUSE and DropoutNet on the RecSys dataset: recall@100, overall training time and recommendation time *per* job seeker (in seconds). DN=DropoutNet, Cold-start=User cold-start.

Recall@	XGBoost	MUSE.0	MUSE.1	MUSE.2
10	26.83	22.88	28.3	**30.1***
20	35.59	31.55	38.0	**40.2***
100	58.88	53.80	61.7	**63.2***
1000	**86.47***	82.13	-	-
Train.	1.83h	7.7h	8.3'	1.25h
Recom.	1.4"	0.0004"	0.018"	0.02"

Table 2: Comparative results of MUSE and XGBoost on the PES dataset: recall@$\{10, 20, 100, 1000\}$, overall training time and recommendation time *per* job seeker (in seconds).

Table 1 reports the recall@100 and computational time of DropoutNet and the MUSE algorithms along two scenarios.[11] The warm start recommendation scenario considers test job ads and job seekers present in the training set, allowing DropoutNet to exploit the latent representation extracted from the decomposition of the interaction matrix, referred to as *I-latent* representation.[12] In warm-start mode, DropoutNet very significantly outperforms all MUSE variants, while MUSE.1 notably improves on MUSE.0. This performance gap is blamed on the fact that MUSE does not use the *I-latent* representation, thus missing any general hint about the interaction matrix and the job market.

In the user-cold scenario, DropoutNet proceeds by aligning the hidden layer representation of the job seekers and job ads (referred to as *S-latent* representation) and the *I-latent* one, enabling the network to retain some general perception of job seekers and job ads w.r.t. the job market.

As could have been expected, the recall@100 in the cold-start scenario is degraded compared to the warm-start one. The gap is very significant for DropoutNet (from 41% to 23%) and much lesser so for MUSE (from 25% to 24% for MUSE.2).

The significant improvement of MUSE.1 compared to MUSE.0 in both scenarios is explained from the fact that MUSE.1 builds upon the pre-selection of the top 1,000 job ads enabled by MUSE.0 (the recall@1,000 of MUSE.0 is 35%). This filter allows for a refined negative sampling in training mode, selecting job ads $\mathbf{y}'$ (Eq. 1) better suited on average to the job seeker $\mathbf{x}$ than random job ads. In inference mode, the filtering of the top 1,000 candidate job ads is key to the low computational cost.

Interestingly, in user-cold start mode, MUSE.1 and DropoutNet have similar performances, and MUSE.2 slightly but statistically outperforms both. A tentative interpretation for this fact is that both MUSE.1 and MUSE.2 exploit the score and rank associated with a pair $(\mathbf{x}, \mathbf{y})$ by MUSE.0: this information expectedly gives some hint into the global structure of the job market, though in the perspective of the job seeker only. Further work will investigate the use of a better exploitation of the MUSE.0 output, e.g. considering also the rank of $\mathbf{x}$ for $\mathbf{y}$ based on MUSE.0$(\mathbf{x}, \mathbf{y})$.

The fact that MUSE.2 improves on MUSE.1 suggests that the RecSys application matrix (gathering only the "impression" interactions) does indeed yield a sufficiently diversified information about the job seekers' preferences compared to the hiring matrix (gathering all other interactions).

5.2 Experimental Results on the *Pôle Emploi* Data

Table 2 reports the recall@$\{10, 20, 100, 1000\}$ and computational time of XGBoost and MUSE on the proprietary *Pôle emploi* dataset.

The main finding is that all MUSE variants but MUSE.0 significantly outperform XGBoost wrt recall@10, 20 and 100, with an inference runtime lesser by two orders of magnitude. These good performances in both terms of recall and runtime are explained from the filter built on the top of the MUSE.0 score: On one hand, the recall@1000 of MUSE.0 is circa 82%, upper bounding by construction the recalls of MUSE.1 and MUSE.2 (though not in a significantly detrimental way). On the other hand, the filter based on the MUSE.0 score contributes to the quality of the learned model, *re* the description of the data and the algorithm itself. At the level of the description of the $(\mathbf{x}, \mathbf{y})$ pairs, the filter enables to consider the expensive $Var(\mathbf{x}, \mathbf{y})$ features (section 3.2, reminding that these features are also provided to XGBoost for a fair comparison). As said, this filter also contributes to a more educated negative sampling, as job ads $\mathbf{y}'$ are now selected among the top 1,000 jobs suited to $\mathbf{x}$.

MUSE.1 significantly improves on MUSE.0 for all recall indicators. It performs on par with the first head of MUSE.2 (also trained to predict the hiring interactions). Note that the second head of MUSE.2 (trained to predict hiring and application interactions alike) is only slightly outperformed by the first head of MUSE.2 regarding its recall on the hiring interactions (recall@10 = 28.4, vs 29.1 for the first head). The key result is that the top head of MUSE.2 (built on the top of the first and second head and trained to predict the hiring interactions) manages to improve on MUSE.1 by about 2 recall levels regarding the hiring interactions. A tentative interpretation for this improvement is that the internal representation (shared by both heads of MUSE.2), referred to as *S-latent* representation, is more representative of the job seekers and job ads, as it leverages a less sparse interaction matrix. Intuitively, the *S-latent* representation can be viewed as a non-linear analogous to the *I-latent* representation (section 5.1). Further work will be devoted to investigate and compare the metrics based on the *I-* and *S-latent* representations, notably depending on the sparsity of the interaction matrix.

[11] The MUSE code source is publicly available at https://gitlab.com/solal.nathan/vadore_ijcai.

[12] As noted by [Volkovs *et al.*, 2017b], taking the scalar product of the latent job seeker and job ad representation even outperforms DropoutNet in warm-start mode (recall@100=42.6%).

	Single module			All modules but one		
R@	M_{geo}	M_{Gal}	M_{sk}	M_{geo}	M_{Gal}	M_{sk}
100	15.43	34.79	4.80	39.97	47.28	51.96

Table 3: MUSE.0: Impact of the three geographical, skills and general modules on the recall@100 through ablation studies. Left: module standalone. Right: MUSE.0 with all modules but one.

The merits of the MUSE.0 architecture are further investigated using ablation studies, aimed to determine the contribution of a standalone module (geographical, skills, general) to the recall performance (Table 3, left). The complementarity of the modules is also examined by removing a single module from the overall architecture (Table 3, right: all modules but one). These results confirm the importance of the geographical module (standalone recall@100 circa 15%; loss in recall@100 circa 14% when omitted). The skill module shows a lesser impact of the skill module (standalone recall@100 circa 4%; loss in recall@100 circa 2% when omitted). More surprising is the impact of the general module (standalone recall@100 circa 34%; loss in recall@100 circa 7% when omitted). Its standalone performance suggests that it contains a larger share of the data information compared to the other modules; on the other hand, the moderate loss suffered when removing the general module suggests that this information is partially redundant with that of the other modules (particularly, the skill module also has access to the occupational profile of job seekers/job ads). Finally, the overall performance of MUSE.0 (recall@100 = 53.8) is close to the sum of the performances of its modules ($15.43 + 34.79 + 4.80 = 55.02$), demonstrating their complementarity.

6 Acceptability and Fairness Study

As said, the social and ethical impacts of recommender systems are increasingly being considered [Ekstrand *et al.*, 2022; Wang *et al.*, 2023; Li *et al.*, 2022], particularly so in the human resources domain. This section first reports on the acceptability study conducted on MUSE, compared to a preference-based system (PBS) inspired from the proprietary *Pôle emploi* system.[13] The fairness of the MUSE recommender system is thereafter investigated.

6.1 Acceptability: An Online Testing Study

Job recommendation systems only trained from past hires may not provide suitable job recommendation in the perspective of *Pôle emploi*. Recommending a job too far from the job seeker's view of their profile might be considered offensive.[14]

[13]The home-made PBS computes the weighted sum of criteria (e.g., working hours, reservation wage, geographic mobility, type of contract; skills, diploma, languages, experience, driving license), measuring the adequacy between the job seeker's preferences and profile, and the job ad, using the same criteria and weights as the proprietary *Pôle emploi* system. PBS was used as the actual *Pôle emploi* system was not accessible on a large scale for technical reasons.

[14]The issue of People With Disabilities is particularly critical: the type of disability is not documented due to regulation policies.

	Clicks	Global	Adequacy	Hiring
MUSE	**0.48***	5.15	3.27	**3.55***
	(1.01)	(2.84)	(3.02)	**(3.00)**
PBS	0.41	**5.17**	**3.30**	3.39
	(0.89)	**(2.80)**	**(3.08)**	(2.99)

Table 4: Online testing of MUSE.0 and PBS on 20,000 job seekers: Average number of clicks on the proposed job ads; Average score (std deviation) for criteria: Global appreciation, Adequacy to preferences and estimated Hiring chance.

The acceptability of the MUSE.0 recommendations is assessed using an online testing on 20,000 job seekers, uniformly divided into a control group receiving the recommendations of the PBS recommendation system, and a treated group, receiving the MUSE.0 recommendations. The recommendation policies based on PBS and MUSE significantly differ: the top-1 job ad recommended by MUSE is included in the top-10 recommendations of PBS for only circa 15% of the job seekers; it does not appear among the top-100 recommendations of PBS for circa 64% of the job seekers.

The online testing, conducted in March 2022 is organized as follows. The control (respectively, treated) job seekers are proposed the top-10 recommendations of PBS (resp. MUSE.0) and they must assess the top-2 job ads along 3 criteria: global evaluation, adequacy to their preferences, estimated chances of hiring. The platform also monitors their clicks on the job ads. The final response rate is 14%; 4% of the job seekers clicked on at least one job ad.

The feedback of the job seekers on the job ads proposed by both systems is very similar (Table 4), except for the number of clicks and the estimated hiring chance, where MUSE.0 slightly but statistically significantly outperforms PBS. Informally, when the job ads proposed by PBS or MUSE.0 are judged negatively, they get the same comments (e.g., "too far"; "I am not interested in this type of job anymore"), suggesting that the acceptability of MUSE.0 recommendations is similar to that of PBS.

6.2 Bias Analysis

As models trained from data might reproduce and increase the prejudices and discriminations involved in human practices, a specific analysis of the recommendation biases is conducted. The analysis of MUSE focuses on the algorithm fairness with respect to gender, known to be an important factor in the study of labor market inequalities.

A first gender-related gap concerns the performances: the recall@10 is respectively 31% for women and 29% for men, the difference being statistically significant (noting that the training data involves 47% men and 53% women). A tentative interpretation for this difference is that the women's labor market behavior might be more focused (hence easier to model) due to the high weight put on geographic distance in labor market trade-offs [Le Barbanchon *et al.*, 2020].

Of course, the gender-related gap might also be inspected w.r.t. any other feature R (wage, distance in kilometers of workplace to job seeker's zip code, whether the job is an executive position, whether the contract is defined for an indefinite duration, number of hours per week, the share of women

in the job sector and the adequacy to the job seeker's preferences). Noting R the specific outcome and G in $\{0, 1\}$ the gender, the gender-related gap is measured using the potential outcome framework [Imbens and Rubin, 2015]:

$$\tau(R) = \mathbb{E}[R^*(1) - R^*(0)] \quad (3)$$

where $R^*(1)$ and $R^*(0)$ are the potential recommendations for women and men. As women and men might have different profiles and preferences, the gender gap is controlled w.r.t. the input search parameters $Z \in \mathcal{Z}$, where Z ranges in: job type, experience, diplomas, desired part time work, desired contract type and wage, qualification level of desired position, accepted mobility, and geographic location. The estimation of $\tau(R)$ relies on the mainstream assumptions of unconfoundedness ($\{R^*(0), R^*(1)\} \perp G \,|Z$) and overlapping support: ($p(Z) := \mathbb{P}(G = g|Z = z) > 0 \quad \forall g \in \{0, 1\}, z \in \mathcal{Z}$). The counterfactual R^* is estimated as:

$$R^*(G) = \mu_0(Z) + \tau G + \varepsilon, \quad (4)$$

where μ_0 models the dependency of R wrt to Z only, and ε is a noise variable independent of G and Z. As the input search parameters Z might depend on the gender, the estimation is corrected using the propensity score $p(Z)$, estimating the probability of $G = 1$ depending on Z. The doubly robust machine learning (DML) estimation method [Chernozhukov *et al.*, 2018] is used to avoid the statistical issues due to learning μ_0 and τ from the same data.

A first difficulty is that the overlapping support assumption does not hold: quite a few job sectors are predominantly occupied by either men or women. For this reason, only job seekers with propensity $p(Z)$ (estimated as the calibrated prediction of G from Z) in [0.05, 0.95] are considered, leaving out circa one third of the job seekers.

The analysis reported in Table 5 compares the gender-related gap estimated from Eq. 4 and associated with the recommendations (τ_{Muse}), the hirings (τ_{Hire}) and the applications (τ_{App}). Eventually, in order to estimate whether the difference between τ_{Muse} and τ_{Hire} (respectively, τ_{Muse} and τ_{App}) is statistically significant, model τ_{DiffH} (resp. τ_{DiffA}) is likewise estimated from Eq. 4, considering the difference between the recommendation and the hiring (resp. the application). Table 5 (column δ) first reports the naive difference of the expectations ($\delta(R) = E[R|G = 1] - E[R|G = 0]$), stating that women are recommended less paid jobs (by .7 point), closer to their location, more often of definite duration and part time (by 17 points), less often in predominantly male job sectors (by 42 points), and requiring one month less in experience. When controlling for Z (column τ_{Muse}), those gaps are reduced. Table 5 (columns τ_{Muse} and τ_{Hire}) also shows that the key gender gaps (related with wages, executive positions, indefinite duration contracts, or adequacy of job ads to job seekers' criteria) are similar for the recommendations and the actual hiring. Conditional gaps in terms of occupation segregation, full time contracts, hours worked are even reduced compared to the actual hirings. While the biases measured on hiring data reflect both job seeker and recruiter preferences, the biases measured on applications might better reflect the job seekers preferences. Most interestingly, the recommendations appear closer to the applications than the hirings ($\tau_{DiffA} < \tau_{DiffH}$).

	δ	τ_M	τ_{Hire} / τ_{DifH}	τ_{App} / τ_{DifA}
LW	-7*	-6*	-5* / -1	-6* / -1
Di	-347*	-182	-2,628* / 2,258*	-1,942* / 1,743*
XP	-1	-4	-6* / 2	-4* / 2
ID	-55*	-33*	-31* / 0	-39* / 7
Ex	-997*	519*	-133 / 551*	105 / 242
MS	-462*	-116*	-17* / 51*	-172* / 151*
FT	-177*	-66*	-101* / 34*	-92* / 33*
HW	-2,547*	-825*	-1,518* / 648*	-1,366* / 597*
Ad	-31*	-23*	-27* / 3	-3* / 4

Table 5: Gender-related gaps ($\times 10^{-3}$) measured from empirical differences in recommendations (δ); counterfactual differences (Eq. 4) associated with recommendations ($\tau_M = \tau_{Muse}$), hirings (τ_{Hire}) and applications (τ_{App}) and their differences τ_{DifH} and τ_{DifA} (see text) w.r.t. LogWage (LW); Distance (Di); Executive position (XP); Indefinite duration (ID); Experience (Ex); Predominently male sector (MS); Full time (FT); Hours work (HW); Adequacy (Ad).

7 Conclusion and Perspectives

This paper tackles the specifics of e-recruitment, as generally faced by national public employment services, involving a vast majority of low-qualified job seekers and an extremely sparse interaction matrix. The presented MUSE approach addresses both challenges through a two-tier architecture, supporting the fast response of the system in inference mode, and allowing the use of informative and computationally demanding features thanks to filtering out most job ads but the most relevant ones to a job seeker. The online testing with real users shows that the acceptability of the recommendations is similar to that of (a home-made implementation of) the proprietary *Pôle emploi* system. The standard recall performance indicators show that MUSE favorably compares with the state of the art with a gain of 2 orders of magnitude in inference time on XGBoost. Last, but not least, the recommendation does not increase the unfairness of the job market, with respect to the gender gap.

A first perspective for further research is to provide the system with a enhanced perception of the general job market and the positioning of the job seeker and job ad within this market, typically using the rank of the job seeker w.r.t. a job add. Refined neuronal architectures, e.g. including a head in charge of predicting the popularity of a particular job ad, will be designed and exploited in order for instance to examine whether and to which extent the perception of the job ad popularity can impact the application decision.

Another perspective is to delve deeper into the trade-off between fairness objectives, such as gender-neutral recommendations using adversarial strategies [Edwards and Storkey, 2016], and recommendation quality, as measured by recall. An intriguing question arises as to whether this trade-off can be customized on an individual basis.

Finally, a promising research perspective is to examine how large language models can be leveraged to better take into account the textual information involved in job ads and in job seekers' resumes, going beyond the domain ontologies relating the job sectors and the skills.

Acknowledgements

We warmly thank C. Vessereau, S. Robidou and P. Beurnier from *Pôle emploi* for making this research possible and granting access to the proprietary data. First author was funded on a grant from the DataIA Institute, Saclay.

References

[Abel *et al.*, 2016] Fabian Abel, András Benczúr, Daniel Kohlsdorf, Martha Larson, and Róbert Pálovics. Recsys challenge 2016: Job recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 425–426. ACM, 2016.

[Abel *et al.*, 2017] Fabian Abel, Yashar Deldjoo, Mehdi Elahi, and Daniel Kohlsdorf. Recsys challenge 2017: Offline and online evaluation. In *Proceedings of the 11th ACM Conference on Recommender Systems, RecSys 2017, August 27-31, 2017*, pages 372–373. ACM, 2017.

[Biancofiore *et al.*, 2021] Giovanni Maria Biancofiore, Tommaso Di Noia, Eugenio Di Sciascio, Fedelucio Narducci, and Paolo Pastore. Guapp: Enhancing job recommendations with knowledge graphs. In *Proceedings of the 11th Italian Information Retrieval Workshop 2021*, volume 2947 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[Borisyuk *et al.*, 2016] Fedor Borisyuk, Krishnaram Kenthapadi, David Stein, and Bo Zhao. Casmos: A framework for learning candidate selection models over structured queries and documents. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 441–450. ACM, 2016.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM, aug 2016.

[Chernozhukov *et al.*, 2018] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[Edwards and Storkey, 2016] Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings*, 2016.

[Ekstrand *et al.*, 2022] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1-2):1–177, 2022.

[Fernandez and Gallardo-Gallardo, 2020] Vicenc Fernandez and Eva Gallardo-Gallardo. Tackling the HR digitalization challenge: key factors and barriers to HR analytics adoption. *Competitiveness Review: An International Business Journal*, 31:162–187, 07 2020.

[Freire and de Castro, 2021] Mauricio Noris Freire and Leandro Nunes de Castro. e-recruitment recommender systems: a systematic review. *Knowledge and Information Systems*, 63:1–20, 2021.

[Geyik *et al.*, 2018] Sahin Cem Geyik, Qi Guo, Bo Hu, Cagri Ozcaglar, Ketan Thakkar, Xianren Wu, and Krishnaram Kenthapadi. Talent search and recommendation systems at linkedin: Practical challenges and lessons learned. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1353–1354. ACM, 2018.

[Imbens and Rubin, 2015] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[Islam *et al.*, 2021] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of the Web Conference 2021*, pages 3779–3790. ACM, April 2021.

[Kenthapadi *et al.*, 2017] Krishnaram Kenthapadi, Benjamin Le, and Ganesh Venkataraman. Personalized job recommendation system at linkedin: Practical challenges and lessons learned. In *Proceedings of the 11th ACM Conference on Recommender Systems, RecSys 2017, August 27-31, 2017*, pages 346–347. ACM, 2017.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[Le Barbanchon *et al.*, 2020] Thomas Le Barbanchon, Roland Rathelot, and Alexandra Roulet. Gender Differences in Job Search: Trading off Commute against Wage. *The Quarterly Journal of Economics*, 136(1):381–426, 10 2020.

[Li *et al.*, 2016] Jia Li, Dhruv Arya, Viet Ha-Thuc, and Shakti Sinha. How to get them a dream job?: Entity-aware features for personalized job search ranking. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016*, pages 501–510. ACM, 2016.

[Li *et al.*, 2021] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1054–1063. ACM, jul 2021.

[Li *et al.*, 2022] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: A survey. *arXiv*, abs/2205.13619, 2022.

[Lian *et al.*, 2014] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *The 20th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, pages 831–840. ACM, 2014.

[Mashayekhi *et al.*, 2022] Yoosof Mashayekhi, Nan Li, Bo Kang, Jefrey Lijffijt, and Tijl De Bie. A challenge-based survey of e-recruitment recommendation systems. *arXiv*, abs/2209.05112, 2022.

[Ozcaglar *et al.*, 2019] Cagri Ozcaglar, Sahin Cem Geyik, Brian Schmitz, Prakhar Sharma, Alex Shelkovnykov, Yiming Ma, and Erik Buchanan. Entity personalized talent search models with tree interaction features. In *The World Wide Web Conference, WWW 2019, May 13-17, 2019*, pages 3116–3122. ACM, 2019.

[Ramanath *et al.*, 2018] Rohan Ramanath, Hakan Inan, Gungor Polatkan, Bo Hu, Qi Guo, Cagri Ozcaglar, Xianren Wu, Krishnaram Kenthapadi, and Sahin Cem Geyik. Towards deep and representation learning for talent search at linkedin. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, October 22-26, 2018*, pages 2253–2261. ACM, 2018.

[Rus *et al.*, 2022] Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H. Schoenmacker. Closing the gender wage gap: Adversarial fairness in job recommendation. *arXiv*, abs/2209.09592, 2022.

[Russakovsky *et al.*, 2014] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2014.

[Shalaby *et al.*, 2017] Walid Ahmed Fouad Shalaby, BahaaEddin AlAila, Mohammed Korayem, Layla Pournajaf, Khalifeh AlJadda, Shannon P. Quinn, and Wlodek Zadrozny. Help me find a job: A graph-based approach for job recommendation at scale. *2017 IEEE International Conference on Big Data (Big Data)*, pages 1544–1553, 2017.

[Sisodia *et al.*, 2017] Dilip Singh Sisodia, Somdutta Vishwakarma, and Abinash Pujahari. Evaluation of machine learning models for employee churn prediction. In *2017 International Conference on Inventive Computing and Informatics (ICICI)*, pages 1016–1020, 2017.

[Volkovs *et al.*, 2017a] Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. Content-based neighbor models for cold start in recommender systems. In *Proceedings of the Recommender Systems Challenge 2017 - RecSys Challenge 17*. ACM Press, 2017.

[Volkovs *et al.*, 2017b] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. DropoutNet: Addressing cold start in recommender systems. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Wang *et al.*, 2023] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 41(3), feb 2023.

[WCC, 2023] WCC. Elise matching: Smart search & match. *https://www.wcc-group.com/employment/products/elise-job-matching-search-and-match/*, visited on the 1st of March 2023.

[Weinberger and Saul, 2009] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, jun 2009.

[Xiao *et al.*, 2016] Wenming Xiao, Xiao Xu, Kang Liang, Junkang Mao, and Jun Wang. Job recommendation with hawkes process: an effective solution for recsys challenge 2016. In *Proceedings of the 2016 Recommender Systems Challenge, RecSys Challenge 2016, September 15, 2016*, pages 11:1–11:4. ACM, 2016.

[Yadav *et al.*, 2018] Sandeep Yadav, Aman Jain, and Deepti Singh. Early prediction of employee attrition using data mining techniques. In *2018 IEEE 8th International Advance Computing Conference (IACC)*, pages 349–354, 2018.

[Yagci and Gurgen, 2017] Murat Yagci and Fikret Gurgen. A ranker ensemble for multi-objective job recommendation in an item cold start setting. In *Proceedings of the Recommender Systems Challenge 2017*, RecSys Challenge '17. ACM, 2017.

[Zhang *et al.*, 2016] XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, and Deepak Agarwal. Glmix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016*, pages 363–372. ACM, 2016.

[Zhao *et al.*, 2021] Jing Zhao, Jingya Wang, Madhav Sigdel, Bopeng Zhang, Phuong Hoang, Mengshu Liu, and Mohammed Korayem. Embedding-based recommender system for job to candidate matching on scale. *CoRR*, abs/2107.00221, 2021.