Recommending the Most Effective Intervention to Improve Employment for Job Seekers with Disability

Ha Xuan Tran[1], Thuc Duy Le[1], Jiuyong Li[1], Lin Liu[1], Jixue Liu[1], Yanchang Zhao[2], Tony Waters[3]
[1]University of South Australia, Adelaide, [2]Data61, CSIRO, Canberra, [3]Maxima Training Group (Aust) Ltd., Adelaide
ha.tran@mymail.unisa.edu.au,{thuc.le,jiuyong.li,lin.liu,jixue.liu}@unisa.edu.au
yanchang.zhao@csiro.au,tony.waters@maxima.com.au

ABSTRACT

In Disability Employment Services (DES), a growing problem is recommending to disabled job seekers which skill should be upgraded and the best level for upgrading this skill to increase their employment potential most. This problem involves counterfactual reasoning to infer causal effect of factors on employment status to recommend the most effective intervention. Related methods cannot solve our problem adequately since they are developed for non-counterfactual challenges, for binary causal factors, or for randomized trials. In this paper, we present a causality-based method to tackle the problem. The method includes two stages where causal factors of employment status are first detected from data. We then combine a counterfactual reasoning framework with a machine learning approach to build an interpretable model for generating personalized recommendations. Experiments on both synthetic datasets and a real case study from a DES provider show consistent promising performance of improving employability of disabled job seekers. Results from the case study disclose effective factors and their best levels for intervention to increase employability. The most effective intervention varies among job seekers. Our model can separate job seekers by degree of employability increase. This is helpful for DES providers to allocate resources for employment assistance. Moreover, causal interpretability makes our recommendations actionable in DES business practice.

CCS CONCEPTS

• **Information systems** → *Decision support systems.*

KEYWORDS

Disability employment, intervention recommendation.

ACM Reference Format:
Ha Xuan Tran[1], Thuc Duy Le[1], Jiuyong Li[1], Lin Liu[1], Jixue Liu[1], Yanchang Zhao[2], Tony Waters[3]. 2021. Recommending the Most Effective Intervention to Improve Employment for Job Seekers with Disability. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3447548.3467095

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467095

Table 1: Rules to answer counterfactual questions

Skill	Characteristics	Empl. increase	Best Int. level	#
Education	Age<50, Gender=Female	0.2	Bachelor	1
	Age<50, Gender=Male	0.3	Trade	2
	Age>=50	0.1	Diploma	3
Computer skill	Disability=Phys.	0.1	Intermediate	4
	Disability=Psyc., Age<50	0.25	Intermediate	5
	Disability=Psyc., Age>=50	0.15	Basic	6
Driving skill	Disability=Phys.	0.2	S. licence	7
	Disability=Psyc.	0.1	D. licence	8

Phys: physical; Psyc: psychological; S: Standard; D: Disability. Int: intervention; Empl: employment chance.

1 INTRODUCTION

People with disability are under-represented in the labour market and tend to experience long-term unemployment [17, 21]. DES providers are major bodies providing expert consultation for disabled people to improve their employment potential. A growing problem confronting providers is which skills job seekers should upgrade to increase their employability most. Given a job seeker's profile, many skills can be considered for upgrading, such as computer skill, language skill or education. The key challenge is that *which skill* should be recommended for upgrading and *the best level for upgrading this skill* to achieve maximum increase in employability. It is infeasible for a job seeker to improve all skills.

We use an example to illustrate the problem. A male job seeker with physical disability wants to increase his employability. He is 35 years old, has a high school certificate but no driver's license. His computer skill is fundamental. The first question is upgrading which skill is the most effective to *increase* his employment chance. Assume that he is advised to upgrade his education. The second question is which education level, among Trade, Diploma,... and PhD, is the best level for upgrading to increase his employability.

The problem involves counterfactual reasoning to answer what-if questions, e.g. how much (what) employment chance would increase if the job seeker intervened (or upgraded) education? The increase in employment chance is the difference between the counterfactual employment chance if this person intervened a factor and the factual employment chance before intervention. A factor and its best intervention level that would create maximum increase in employment chance are recommended to the job seeker.

Table 1 shows an example of rules that can be used to answer the above counterfactual question for three factors. For instance, rule 4 shows that if the job seeker intervened computer skill, his employment chance would increase by 0.1. Among three factors, upgrading education is the most effective intervention to increase employment chance (rule 2, 4, 7). Trade is the education level that he needs to obtain in order to achieve the best employability increase.

The problem can be formulated as recommending the most effective intervention to increase employability. Both a relevant factor and its best intervention level are to be recommended. Solution requires interpretability since intervention entails major decisions.

Popular employment recommendation systems cannot solve our problem adequately. They provide candidate recommendations [2, 3, 26], job recommendations [3, 11, 12], or skill recommendations [7, 9, 31]. Problems solved by these systems do not require counterfactual reasoning. Their main goal is to optimize employment status based on association between factors and the outcome [33]. Our problem focuses on *increase* in employment chance. It requires addressing counterfactual questions to produce recommendations.

Other related areas are treatment effect heterogeneity and uplift modelling. Causal effects are estimated for subgroups. Many tree-based methods [4, 6, 14, 25, 30, 33, 38] are developed to achieve interpretability. They work with binary factors or randomized trials. The question about the best intervention levels of continuous/ordinal factors for subgroups using observational data remains open. Most current methods assume known causes in estimation. In our problem, causal factors need to be detected from data.

We solve the problem by a causal method with two stages. First, we detect causal factors of employment status from data. Second, we combine a counterfactual framework with a machine learning approach to build causal trees for estimating causal effects and the best intervention levels. Causal effect measures employability increase created by an intervention. Building a causal tree involves multi-objective optimization. It requires maximizing effect fitness, effect diversity, and intervention level diversity upon splitting.

We apply our method to synthetic datasets and a case study with real data of 4697 Australian disabled jobseekers. Experiments consistently show real promise for improving employability of job seekers. Factors, if intervened, would increase employment chance are motivation, education, and work capacity. The most effective factor with its best intervention level vary among job seekers. Our method can separate job seekers into groups with different degrees of employability increase. This is helpful for DES providers to allocate resources for employment assistance. Our model is causally interpretable as recommendations are based on causal effects estimated by causal trees. Our contributions are summarised as below:

- We present the first study of recommending a factor and its best intervention level to achieve *maximum employability increase* for disabled job seekers. This entails counterfactual reasoning to recommend the most effective intervention.
- We develop a two-stage method with causal interpretability It involves causal learning and multi-objective optimization.
- We apply our method to synthetic data and a case study for performance evaluation and answering domain questions.
- We disclose effective factors for intervention to increase employment chance of job seekers. The most effective factor and its best intervention level vary among job seekers. Practical implications are also introduced for the DES sector.

2 PROBLEM SETUP

2.1 Preliminaries

Let a job seeker be described by a set of features $\mathbf{X} = \{X_1, X_2, ..., X_d\}$ where X_i has continuous or ordinal values. Y, a binary variable, is employment status of the job seeker, with $Y = 1$ for being employed and $Y = 0$ for being unemployed. Variable set $\mathbf{F} \subset \mathbf{X}$ describes manipulable features, e.g. education level. $\mathbf{F}$ is called a set of factors.

Since our problem involves counterfactual reasoning, causation provides a base to cope with this challenge. Below are the main concepts required to frame our problem in the context of causation.

Causal relationship. Two variables are said to have a causal relationship if changes in one variable lead to changes in the other variable. The first variable is called a cause of the second. Normal association between two variables does not have this property [20].

Causal effect. In our problem, the fundamental counterfactual question is how much employment chance would increase if a factor was intervened. Addressing this question is a process of causal inference. Determining causal effect of a factor on employment status provides a quantitative answer to that question.

To define causal effect of factor F on employment staus Y, we adopt the counterfactual framework [24]. It is developed to estimate the effect of a treatment on patients with recovery status $Y \in \{0, 1\}$, $Y = 1$ for being recovered and $Y = 0$ for being unrecovered. Let $T \in \{0, 1\}$ be a treatment indicator for patient i with two potential outcomes $Y_i^{T=1}$ and $Y_i^{T=0}$ for receiving and not receiving treatment P. Individual treatment effect (ITE) of treatment P for this person is defined as $Y_i^{T=1} - Y_i^{T=0}$. Since a person cannot be treated and non-treated at the same time, only one potential outcome is observed. This makes it challenging to directly identify ITE. Instead, ITE is approximated by the expected effect of treatment P for a subgroup that patient i belongs to. It is called conditional average treatment effect ($CATE$), which is defined as

$$\tau = \mathbb{E}[Y_i^{T=1} - Y_i^{T=0} | \mathbf{C} = \mathbf{c}_i], \tag{1}$$

where $\mathbf{C}$ is a variable set describing the subgroup and $\mathbf{C} = \mathbf{X} \setminus \mathbf{F}$.

Assume that factor F is a binary variable where $F = 1$ means intervening the factor and $F = 0$ means not intervening it. We map treatment P to factor F, and recovery status to employment status. $CATE$, thereby, approximately represents causal effect of factor F on employment status Y for an individual. It measures how much the employment chance of this person would increase by intervening factor F. The larger τ, the higher increase in employment chance.

Since F is a continuous or ordinal variable, we redefine treatment indicator T to be the amount of intervening factor F for each job seeker. Let θ be a threshold that value of F needs to reach. We convert a problem of continuous/ordinal treatment to a problem of binary treatment with the concept of an intervention threshold. $CATE$ of factor F for a given intervention threshold θ is defined as

$$\tau(\theta) = \mathbb{E}[Y_i^{F>=\theta} - Y_i^{F<\theta} | \mathbf{C} = \mathbf{c}_i]. \tag{2}$$

2.2 Objectives

We have three objectives that need to be achieved in our problem.

1) Detecting factors for intervention. We want to identify causes of employment status Y from $\mathbf{F}$. It is unreasonable to intervene a factor that has no effect on the outcome. A causal model for each detected factor is built to produce recommendations for that factor. The next objectives describe requirements for such a model.

2) Estimating the best intervention level and maximum causal effect. For a causal factor F, the best intervention level is defined as a level required to achieve maximum increase in employment chance. Estimation needs to be performed at subgroup levels for personalized recommendations. Disabled job seekers are partitioned into subgroups based on their characteristics. The best intervention level of factor F for a subgroup is given as

$$\theta^* = \underset{\theta}{\text{argmax}}(\mathbb{E}[Y_i^{F>=\theta} - Y_i^{F<\theta}|\mathbf{C} = \mathbf{c}_i]). \quad (3)$$

It is a value that F needs to reach to achieve a maximum causal effect, $\tau^* = \tau(\theta^*)$. That is maximum employability increase.

The partitioning needs to maximize effect fitness to maximize prediction accuracy of τ^*. We also encourage our model in discovering job seeker heterogeneity by maximizing diversity of τ^* and diversity of θ^* among subgroups. This leads to a problem of multi-objective optimization for partitioning, where effect fitness, diversity of τ^* and diversity of θ^* are simultaneously maximized.

Given τ^* and θ^* of all causal factors, we can recommend to a job seeker the most effective factor to increase employment chance and advise the best intervention levels for recommended factors.

3) Interpretability. Our model requires interpretability. This allows job seekers and providers to understand the recommendation process, giving them confidence in adopting our recommendations.

3 METHOD

Our method includes two major tasks. First, we detect causes of employment status from data. Second, we build causal trees to estimate maximum causal effects of these causes and the best intervention levels to achieve those effects for subgroups of job seekers.

3.1 Identifying causes of employment status

3.1.1 Causal mechanism representation. Factors to be intervened need to be causes of employment status Y. We use graphical causal modelling [23] to infer causal structures from data. Causal relationships are popularly represented by a directed acyclic graph (DAG), which contains only directed edges and no cycles.

Definition 1. Causal graph. Given causal graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V} = \{V_1,..., V_p\}$ is a node set and $\mathbf{E}$ is a edge set. $\mathbf{V}$ represents features and employment status of job seekers, i.e. $\mathbf{V} = \mathbf{X} \cup \{Y\}$. A directed edge $V_\text{i} \rightarrow V_\text{j}$ represents a causal relationship where V_i is called a cause or a parent of V_j, and V_j is called a effect of V_i. We use a DAG to represent causal relationships among variables.

To link a DAG with data distribution, the Markov condition [27] is required: a node in a DAG is conditionally independent of all its non-descendants given its parent set. Two other conditions, faithfulness and causal sufficiency [27], are also assumed to achieve a causal DAG (Details in Supplement - Section 8.1.3).

Figure 1 illustrates a causal DAG where education and motivation are causes of employment status. Their causal effects and best intervention levels are estimated to make recommendations.

Given causal factor F and outcome Y, other variables are called covariates. Covariates may contain confounding variables that influence both F and Y. We need to block confounding effects to

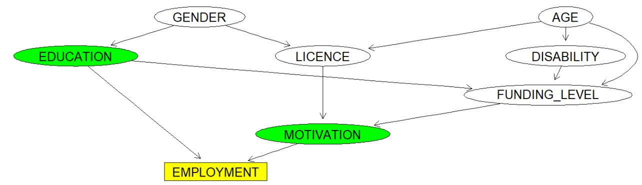

Figure 1: Example of causal DAG.

obtain unbiased estimation for causal effect of F on Y. Causal DAG is a base element of graphical causal inference. From a DAG, we can identify a set of variables for blocking confounding effects.

3.1.2 DAG construction. We use the PC algorithm [27] to learn causal factors of employment status from data. Detected causal relationships, represented by a DAG, are validated by domain experts. If some relationships are unreasonable, we fix them, for example, by removing all incoming edges to the gender node, and fixing an edge between two nodes, such as age causing disability type, which is a known relationship. This process is to encode human belief into a causal DAG, making it consistent with both data and domain knowledge (Details in Supplement - Section 8.1.1).

3.2 Estimating maximum causal effect and the best intervention level

3.2.1 Estimating unbiased causal effect. We follow the counterfactual framework to define average treatment effect (ATE) over the entire population with binary treatment T on outcome Y as $\mathbb{E}[Y^{T=1} - Y^{T=0}]$. To estimate ATE with observational data, we make three assumptions [24]: 1) *treatment assignment of a patient does not affect potential outcomes of other patients,* 2) *observed outcomes of a treatment equal its potential outcomes,* 3) *given covariate set* $\boldsymbol{B}$*, treatment assignment is independent of potential outcomes, i.e.* $T \perp\!\!\!\perp (Y^{T=0}, Y^{T=1})|\mathbf{B}$. Unbiased estimation of ATE is given as [15]:

$$\widehat{ATE} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{e(\mathbf{b}_i)}}{\sum_{i=1}^{n} \frac{T_i}{e(\mathbf{b}_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{(1-e(\mathbf{b}_i))}}{\sum_{i=1}^{n} \frac{(1-T_i)}{(1-e(\mathbf{b}_i))}}, \quad (4)$$

where n is the total number of individuals in the entire population, $e(\mathbf{b}_i) = Pr(T_i = 1|\mathbf{B} = \mathbf{b})$ is a propensity score [15]. It is probability of receiving treatment of a patient described by $\mathbf{b}$. We use logistic regression to estimate propensity scores.

$CATE$ is a special case of ATE, where the treatment effect is estimated at subgroup levels. $CATE$ in Equation 1 can be estimated by Equation 4 over a subgroup. In our problem, we need to estimate $CATE$ in Equation 2 for a continuous/ordinal causal factor F with an intervention threshold θ. It is estimated as

$$\hat{\tau}(\theta) = \frac{\sum_{i=1}^{n} \frac{\mathbb{I}_{\{F_i>=\theta\}} Y_i}{e(\mathbf{b}_i)}}{\sum_{i=1}^{n} \frac{\mathbb{I}_{\{F_i>=\theta\}}}{e(\mathbf{b}_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-\mathbb{I}_{\{F_i>=\theta\}})Y_i}{(1-e(\mathbf{b}_i))}}{\sum_{i=1}^{n} \frac{(1-\mathbb{I}_{\{F_i>=\theta\}})}{(1-e(\mathbf{b}_i))}}, \quad (5)$$

where $\mathbb{I}_{\{F_i>=\theta\}}$ is an indicator function, n is the number of job seekers in a subgroup, and $e(\mathbf{b}_i) = Pr(F_i >= \theta|\mathbf{B} = \mathbf{b})$.

Using Equation 5, the best intervention level and maximum causal effect of F for a subgroup is estimated as $\widehat{\theta^*} = \text{argmax}_\theta(\hat{\tau}(\theta))$, and $\widehat{\tau^*} = \hat{\tau}(\widehat{\theta^*})$ respectively.

Our next task is identifying covariate set $\mathbf{B}$ for blocking confounding effects to achieve unbiased causal effect estimation.

Definition 2. Back-door path. Given DAG $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ and a cause-effect pair ($F \in \mathbf{V}$, $Y \in \mathbf{V}$), back-door path of (F, Y) is a path that indirectly connects F and Y and has no *collider*. Node V_{j} is a collider of a *v-structure* created by three nodes (V_{i}, V_{j}, V_{t}) if V_{j} is a child of both V_{i} and V_{t}, and no edge exists between V_{i} and V_{t}.

We use backdoor criterion [23] to determine $\mathbf{B}$. Since all causal relationships can be learned from data, parents of F are sufficient to block all back-door paths between F and Y, i.e. $\mathbf{B} = PA(F)$.

3.2.2 Building a model for estimating maximum causal effect and the best intervention level. Given factor F, job seekers are partitioned into subgroups whose τ^* and θ^* are estimated for this factor.

Objective function. There are three objectives for partitioning population $\mathcal{X}$. First, we need to maximize effect fitness by minimizing the mean squared error (MSE). Since true causal effect is unobservable, estimation of MSE is challenging. We follow Athey et al. [4] to define effect fitness measure for subgroup $\mathcal{G}_i$ as $N_{\mathcal{G}_i} \cdot \widehat{\tau^*}^2_{\mathcal{G}_i}$, where $N_{\mathcal{G}_i}$ is the number of samples in the subgroup. We want to maximize the sum of effect fitness across partitions. Second, we want to encourage effect diversity among subgroups, i.e. $max(\widehat{\tau^*}_{\mathcal{G}_i} - \widehat{\tau^*}_{\mathcal{G}_j})^2, i \neq j$. Third, we want to maximize intervention level diversity, i.e. $max(\widehat{\theta^*}_{\mathcal{G}_i} - \widehat{\theta^*}_{\mathcal{G}_j})^2, i \neq j$. These objectives are formulated as the following multi-objective optimization problem.

$$\underset{\mathcal{G}_1,\ldots,\mathcal{G}_K}{\operatorname{argmax}} \Big(\sum_{i=1}^{K} N_{\mathcal{G}_i} \cdot \widehat{\tau^*}^2_{\mathcal{G}_i}, \sum_{i=1,j=1}^{K} (\widehat{\tau^*}_{\mathcal{G}_i} - \widehat{\tau^*}_{\mathcal{G}_j})^2, \sum_{i=1,j=1}^{K} (\widehat{\theta^*}_{\mathcal{G}_i} - \widehat{\theta^*}_{\mathcal{G}_j})^2 \Big), \tag{6}$$

where $i \neq j$, $\cup_i^K \mathcal{G}_i = \mathcal{X}$ and $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$.

Since the space to search for subgroups with multi-objective optimization is large, we develop a tree-based method, called Maximum Causal Tree (MCT), for efficient search and interpretability. Each tree is built to estimate maximum causal effect τ^* and the best intervention level θ^* for a factor. We take a greedy approach to build a tree, where an objective function is maximized at each node split. Given a node to be split into child nodes L and R with causal effects $\widehat{\tau^*}_L, \widehat{\tau^*_R}$ and intervention levels $\widehat{\theta^*}_L, \widehat{\theta^*}_R$, a selected split needs to maximize the following objective function.

$$\underset{L,R}{\operatorname{argmax}} \Big((N_L \cdot \widehat{\tau^*}^2_L + N_R \cdot \widehat{\tau^*}^2_R), (\widehat{\tau^*}_L - \widehat{\tau^*}_R)^2, (\widehat{\theta^*}_L - \widehat{\theta^*}_R)^2 \Big), \tag{7}$$

where the three objectives are named as *effect fitness*, *effect diversity* and *intervention level diversity* respectively.

Multi-objective optimization. Let $\mathbf{p}_i = (p_{efi}, p_{edi}, p_{ldi})$ indicate a score vector of effect fitness, effect diversity, and intervention level diversity for split i, where $\mathbf{p}_i \in \mathcal{P}$ and $\mathcal{P}$ is the set of score vectors for all split points, $\mathcal{P} \subset \mathbf{R}^3$. The goal is to find an optimal $\mathbf{p}_i$ from $\mathcal{P}$. We use the term *split* for $\mathbf{p}_i$ if the context is clear.

Pareto dominance is a popular mathematical model for multi-objective optimization [10]. Nonetheless, it is not practically applicable to our problem. First, a Pareto set often has a large number of elements, making it prohibitive for optimization. Second, Pareto dominance does not facilitate trade-off among the three objectives. Extended Pareto dominance is proposed to achieve our goal. It is called ϵ-dominance [19].

Definition 3. ϵ-dominance. Score vector $\mathbf{p}_i$ ϵ-dominates $\mathbf{p}_j$ for some $\boldsymbol{\epsilon} = (\epsilon_{ef}, \epsilon_{ed}, \epsilon_{ld})$, i.e. $\mathbf{p}_i \succ_\epsilon \mathbf{p}_j$, iff $(1 + \epsilon_{ef}) \cdot p_{efi} \geq p_{efj} \wedge (1 + \epsilon_{ed}) \cdot p_{edi} \geq p_{edj} \wedge (1 + \epsilon_{ld}) \cdot p_{ldi} \geq p_{ldj}$.

ϵ-dominance allows to specify different dominance magnitudes for different criteria. Intuitively, $\mathbf{p}_j$ is not ϵ-dominated by $\mathbf{p}_i$ if elements of $\mathbf{p}_j$ are at least larger than $\mathbf{p}_i$ by a margin denoted by $\boldsymbol{\epsilon}$.

Definition 4. ϵ-optimal split set. An optimal ϵ-split set $\mathcal{P}^* \subseteq \mathcal{P}$ satisfies 1) $\forall \mathbf{p}^* \in \mathcal{P}^* : \nexists \mathbf{p} \in \mathcal{P}$ s.t. $\mathbf{p} \succ_\epsilon \mathbf{p}^*$, 2) $\forall \mathbf{p} \in \mathcal{P} \backslash \mathcal{P}^* : \exists \mathbf{p}^* \in \mathcal{P}^*$ s.t. $\mathbf{p}^* \succ_\epsilon \mathbf{p}$.

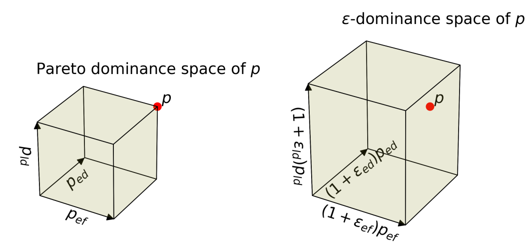

Figure 2: Pareto dominance versus ϵ-dominance.

Figure 2 shows the difference between Pareto dominance and ϵ-dominance. Since the ϵ-dominance space is larger than the Pareto dominance space, $\mathcal{P}^*$ is smaller than the original Pareto set. To further reduce the complexity of searching for an ϵ-optimal split, we follow Laumanns et al. [19] to discretize the search space into cuboids of size $\Big(\lfloor \frac{log p_{ef}}{log(1+\epsilon_{ef})} \rfloor, \lfloor \frac{log p_{ed}}{log(1+\epsilon_{ed})} \rfloor, \lfloor \frac{log p_{ld}}{log(1+\epsilon_{ld})} \rfloor\Big)$.

Algorithm 1 is designed for maintaining ϵ-optimal split set $\mathcal{P}^*$. ϵ-dominance is applied to cuboids to achieve a non ϵ-dominated cuboids. Line 1 means existing splits are removed from optimal set $\mathcal{P}^*$ if cuboids containing these splits are ϵ-dominated by the cuboid of new split $\mathbf{p}$. Line 4-5 means if the cuboid of $\mathbf{p}$ is not ϵ-dominated by any cuboid from $\mathcal{P}^*$, the new split is added into $\mathcal{P}^*$. Within a cuboid, only one split not ϵ-dominated by others is kept (line 2-3). This guarantees convergence with an ϵ-optimal split set.

Algorithm 1 Maintaining an ϵ-optimal split set

Input: Current ϵ-optimal split set $\mathcal{P}^*$, a new split point $\mathbf{p}$.
Output: Updated ϵ-optimal split set $\mathcal{P}^*$.
1: $\mathcal{P}^* := \mathcal{P}^* \setminus \{\mathbf{p}^* \in \mathcal{P}^* | cuboid(\mathbf{p}, \boldsymbol{\epsilon}) \succ_\epsilon cuboid(\mathbf{p}^*, \boldsymbol{\epsilon})\}$
2: **if** $\exists \mathbf{p}^* \in \mathcal{P}^*$: $cuboid(\mathbf{p}^*, \boldsymbol{\epsilon}) = cuboid(\mathbf{p}, \boldsymbol{\epsilon})$ and $\mathbf{p} \succ_\epsilon \mathbf{p}^*$ **then**
3: $\quad \mathcal{P}^* := \mathcal{P}^* \cup \{\mathbf{p}\} \setminus \{\mathbf{p}^*\}$
4: **else if** $\nexists \mathbf{p}^* \in \mathcal{P}^*$: $cuboid(\mathbf{p}^*, \boldsymbol{\epsilon}) \succ_\epsilon cuboid(\mathbf{p}, \boldsymbol{\epsilon})$ **then**
5: $\quad \mathcal{P}^* := \mathcal{P}^* \cup \{\mathbf{p}\}$
6: **end if**
7: **return** $\mathcal{P}^*$

Tree construction. Algorithm 2 describes construction of a maximum causal tree for factor F. For each splitting, the algorithm determines ϵ-optimal split set $\mathcal{P}^*$ (line 8-13). An optimal split must not reduce the effect fitness (line 14-15). A split in $\mathcal{P}^*$ with the largest fitness effect is selected for splitting (line 16-19). This process stops when the max height is reached or effect fitness declines.

EstimateMaxEffect is a function to estimate a maximum causal effect τ^* and the best intervention level θ^* for F. We use an approximate binning strategy for continuous causal factors to reduce the search space for θ^* (Details in Supplement - Section 8.1.2).

From a causal tree, we can determine causal effect and the best intervention level of a factor estimated for a job seeker. We build a causal tree for every causal factor. The factor that has the highest causal effect, and its best intervention level are recommended to this person for intervention to increase his/her employment chance.

Algorithm 2 Maximum Causal Tree (MCT)

Input: Dataset $\mathbf{D}$, causal factor F, max tree depth h_{max}.
Output: A maximum causal tree.

TreeConstruct(0, $\mathbf{D}$, F)
return
Function TreeConstruct(h, $\mathbf{D}$, F)
if $h >= h_{max}$ **then**
  return
end if
let $\mathcal{P}^* = \emptyset$
for each split point $\mathbf{p}_i \in \mathcal{P}(\mathbf{D})$ **do**
  EstimateMaxEffect(LeftChild($\mathbf{p}_i$, $\mathbf{D}$), F)
  EstimateMaxEffect(RightChild($\mathbf{p}_i$, $\mathbf{D}$), F)
  Compute $(p_{efi}, p_{edi}, p_{ldi})$ for $\mathbf{p}_i$
  Update $\mathcal{P}^*$ using Algorithm 1 with $\mathbf{p}_i$
end for
Compute effect fitness of current node using $N \cdot \widehat{\tau^*}^2$
$\mathbf{S} = \{\mathbf{p}^* \in \mathcal{P}^* | p_{ef} \geq$ effect fitness of current node $\}$
if $\mathbf{S} \neq \emptyset$ **then**
  $SelectedSplit = \text{argmax}_{\mathbf{s} \in \mathbf{S}}(p_{ef})$
  TreeConstruct($h + 1$, LeftChild($SelectedSplit$, $\mathbf{D}$), F)
  TreeConstruct($h + 1$, RightChild($SelectedSplit$, $\mathbf{D}$), F)
end if
End

We evaluate our method with synthetic data and a case study with a real-world dataset provided by an Australian DES provider. Results from the case study are used to verify model interpretability, answer domain questions and provide practical implications.

4 SIMULATIONS

4.1 Evaluation method

Our goal is to evaluate how interventions on a given causal factor would increase employment chance of job seekers. Data are created in four scenarios with causal factor F, employment status Y and attributes $\mathbf{X}$. $\mathbf{X}$ follow normal distributions. Distribution of F, which follows truncated normal or Poisson distribution, depends on $\mathbf{X}$ and interaction among attributes. Y follows Bernoulli distribution whose parameters depends on F and $\mathbf{X}$ with different complexity levels (Details in Supplement - Section 8.2).

Baselines. Since our problem involves causation, we compare our method with causation-based methods. However, no existing method based on observational data focuses on recommending intervention levels for subgroups. We use methods designed for binary interventions as baselines. In this way, factor F in a dataset is binarized using breakpoints. A derived binary factor F has value of 0 if its original value is smaller than a breakpoint. Otherwise, the binary factor has a value of 1. With 50 breakpoints, we have 50 datasets derived from each original dataset.

To be consistent with the interpretability objective, we compare with four recent tree-based methods designed for estimating heterogeneous treatment effects for subgroups:

Causal tree (CT) [4]. CT selects a split based on effect fitness criterion. It proposes an honest strategy for unbiased estimation.

Transformed Outcome Tree (TOT) [5]. TOT is based on the idea of using standard regression tree methods to estimate treatment effect with a transformed outcome variable as a target.

Squared T-Statistic Tree (ST) [29]. ST selects a split with the largest difference in the average treatment effect in two leaves.

Fit-based Tree (FT) [37]. The tree is constructed using the CART algorithm. Treatment effect is estimated according to Equation 4.

Evaluation metrics. We use two criteria for evaluating performance. First, we want to know if job seekers followed recommendations given by individual methods, which method would help to increase their employment chance most. Job seeker i is said to follow our recommendations if he/she intervened factor F to a recommended level. For binary interventions, job seeker i followed recommendations by the baselines if binary factor F has the value of 1 for this factor. To evaluate this criterion, we use Area Under Uplift Curve (AUUC) [13] with a modification to the curve. Job seekers are sorted by causal effect in descending order and partitioned into percentile segments, e.g. top 10%. The modified uplift curve is defined as $MUC(\phi) = (\bar{Y}_{R\phi} - \bar{Y}_{NR\phi}) \cdot p \cdot 100$, where ϕ is a segment of top ϕ%, $\bar{Y}_{R\phi}$ is the average of observed employment rate among those following recommendations, $\bar{Y}_{NR\phi}$ is the average observed employment rate among those not following recommendations, and p is the proportion of job seekers in segment ϕ. The larger AUUC, the higher employability increase, hence the better method.

Second, we want to evaluate how our model discovers job seeker heterogeneity, where job seekers can be separated into groups with different degrees of employability increase. To evaluate this criterion, job seekers are also ranked by causal effect in descending order and partitioned into percentile segments. We use the Kendall correlation coefficient and Spearman correlation coefficient to evaluate consistency between orderings by causal effect and by observed employability increase among segments. Observed employability increase of segment ϕ is given as $(\bar{Y}_{R\phi} - \bar{Y}_{NR\phi})$. A better method should have larger values for both coefficients, indicating better accuracy in discovering heterogeneous employability increases. Since each baseline is run with many derived datasets, the derived dataset with the largest with AUUC is used to compute the coefficients.

4.2 Model performance

4.2.1 Employability improvement. Figure 3 shows results of AUUC for five methods. In all scenarios, AUUC of MCT is better than the best values of the baselines and far better than their average values, i.e. blue diamonds. This indicates that intervention levels recommended by MCT would create greater employability increase than the baselines. Interestingly, in scenarios 3 and 4 where the numbers of attributes are more than double the numbers in scenarios 1 and 2, MCT performs much better than the average of baseline methods.

Table 2: Kendall correlation coefficient

Scenario	MCT	CT	TOT	ST	FT
1	**1.00**	0.33	0.82	0.47	0.78
2	**1.00**	0.24	0.20	0.11	0.38
3	**1.00**	0.82	0.16	0.69	0.82
4	**1.00**	0.87	0.02	0.82	0.96

Table 3: Spearman correlation coefficient

Scenario	MCT	CT	TOT	ST	FT
1	**1.00**	0.41	0.92	0.53	0.89
2	**1.00**	0.31	0.20	0.16	0.47
3	**1.00**	0.94	0.24	0.78	0.90
4	**1.00**	0.95	0.10	0.92	0.99

The baselines need to run with all derived datasets to find a good intervention level. This makes it prohibitive for practical applications with continuous causal factors.

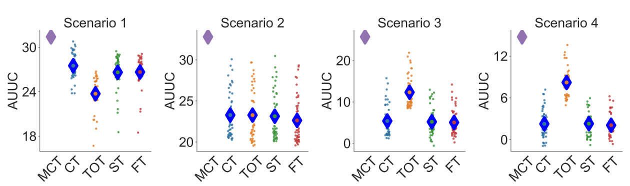

Figure 3: In one scenario, each baseline is run with 50 derived datasets, whereas MCT is run with an original dataset.

4.2.2 Heterogeneity discovery. Table 2 and table 3 show that MCT discovers heterogeneous causal effects with higher accuracy than all baselines. Both measurements indicate our model achieves the highest consistency between orderings by causal effect and by employability increase, among five methods. MCT performs best in separating job seekers into different groups with different levels of employability improvement based on causal effects.

5 A CASE STUDY OF AUSTRALIAN DISABILITY EMPLOYMENT

5.1 Preliminaries

Our industry partner, Maxima Training Group, provides us data extracted from their application platforms to carry out a case study. Below are three objectives we want to achieve for this case study.

1) To evaluate model performance with real data. Evaluation metrics are the same as the previous experiments, but causal factors are detected from data. All methods use these causal factors for building their models. The goal is to recommend the most effective intervention for increasing employment chance of job seekers.

2) To verify interpretability. First, we want to verify whether detected causes for employment status of disabled job seekers are reasonable. Second, our causal trees for producing recommendations are verified whether the trees are interpretable to humans.

Table 4: Attributes of disabled job seekers

Attribute	**Description**
Age	Age of disabled job seekers.
Gender	Gender of disabled job seekers.
Education	Education level of disabled job seekers.
Disability	Disability type of disabled job seekers.
E-communication	Ability to communicate by email.
Licence	Information about driving licences.
Work capacity	Weekly hours job seekers can work.
Motivation	Motivation level of disabled job seekers.
Allowance	Welfare allowance types.
Program	Disability employment support programs.
Funding level	Fee level, depending on employment barrier.
Employment	Employment status.

3) To answer domain questions. Maxima wants to know: a) whether our method would help to increase employability of disabled job seekers; b) all factors are recommended for interventions. Although the most effective factor may vary among job seekers, realizing all factors to be intervened helps DES providers prepare plans for upskilling assistance; c) whether job seekers should upgrade causal factors to the highest levels to achieve the best employability improvement; and d) how to determine degree of employability increase for a job seeker if he/she adopted our recommendations. This is helpful for providers to allocate resources for supporting job seekers in improving their employment chance.

5.2 Dataset

The dataset records information of 4697 Australian disabled job seekers with over 40 attributes. Table 4 shows 11 attributes and an outcome for constructing our model. Causal factors are detected in the first stage of our method. Identified causes are used to build causal trees in the second stage. To feed into the baselines, causal factors are binarized using breakpoints advised by our industry partner. We have 80 ways of binarizing causal factors, and this produces 80 derived datasets. Each baseline is run on these datasets to compare with our method, which is run on the original dataset.

5.3 Model performance

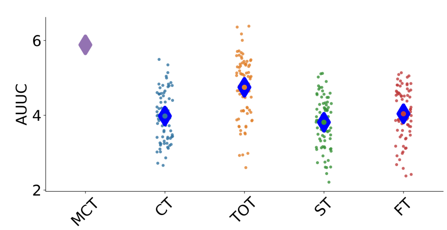

Figure 4: Each baseline is run with 80 derived datasets, whereas MCT is run with an original employment dataset.

Table 5: Kendall and Spearman correlation coefficients

Method	Kendall coefficient	Spearman coefficient
MCT	**1.00**	**1.00**
CT	0.60	0.76
TOT	-0.42	-0.62
ST	-0.56	-0.61
FT	0.47	0.54

5.3.1 Employability improvement. Figure 4 shows MCT reaches the top performance of the baselines that are run on 80 derived datasets to search for the most effective intervention. AUUC of MCT is significantly larger than the average values of baseline methods, i.e. blue diamonds. Our model would create higher employability increase for job seekers if they adopted our recommendations.

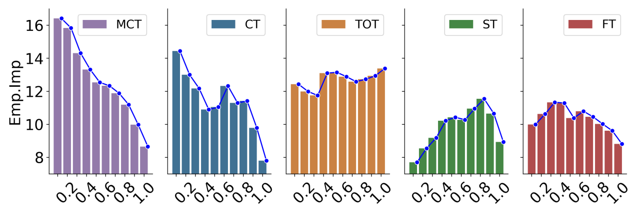

Figure 5: Observed employability increase of top groups.

5.3.2 Heterogeneity discovery. Table 5 shows MCT has better accuracy of discovering the heterogeneity of disabled job seekers. Both Kendall and Spearman coefficients indicate MCT achieves the highest consistency between ranking job seekers by causal effect and ranking by employability increase.

Good capture of heterogeneity is also indicated by diversity of employability increase. To evaluate this criterion, job seekers are ranked by causal effect in descending order, and partitioned into percentile segments, e.g. top 10%. Observed employability increase of segment ϕ is given as $(\bar{Y}_{R\phi} - \bar{Y}_{NR\phi}) \cdot 100$. For each baseline, a derived dataset with the best AUUC is used for comparison.

Figure 5 shows that employability increase of MCT is the largest with the top 10% group and steeply declines with other groups. This means that MCT captures heterogeneity well for accuracy, indicated by a consistently declining trendline, and for diversity of employability increase, indicated by decreased amounts between two consecutive groups. The baselines have inconsistently declining trendlines and/or lower diversity of employability increase.

5.4 Model interpretability

5.4.1 Causal factors of employment status. Figure 6 presents a causal DAG inferred from the data. It shows that the employment status of disabled job seekers is affected by three factors: motivation, education and work capacity. This is reasonable since motivation is believed a strong predictor for participating in the workforce for people with disability [1]. Education is known to have a relationship with the employment rate of disabled job seekers [8]. Our industry partner also confirms that work capacity of disabled candidates is an important criterion for many jobs.

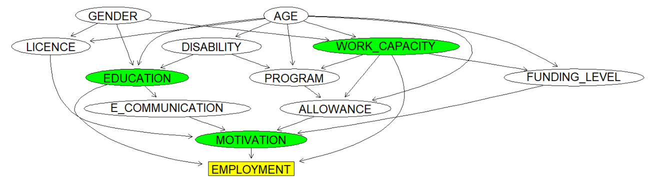

Figure 6: Causal relationships inferred from the data.

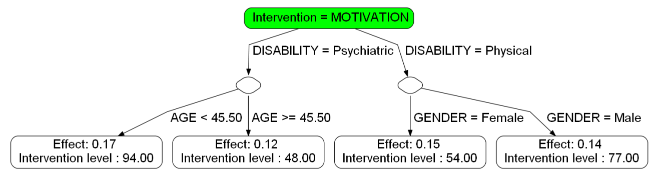

Figure 7: Causal tree for motivation intervention.

5.4.2 Causal trees for recommendation. The causal tree in Figure 7 shows motivation has positive effects on employment status of all subgroups. Disabled people often face barriers to employment. They need to be motivated for work to overcome such barriers.

Among four subgroups, motivation has the smallest effect on old job seekers with psychiatric disability. These people have more disadvantages than others as they suffer more discrimination [17]. This explains why their employment chance would increase less than other job seekers when they improved their motivation.

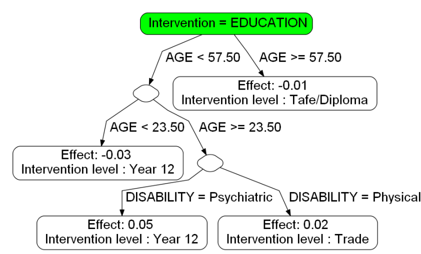

Figure 8: Causal tree for education intervention.

Figure 8 shows that education has smaller effects on employment status than motivation. Disabled people often involve non-professional jobs, which do not require high education. Upgrading education does not help to increase employability much. Improving education even reduces employability of old and immature people. The case of old people can be explained with overqualification. The reason for young people can be increased motivation for further study instead of working after graduating from high school.

Figure 9 shows that upgrading work capacity would increase employment chance. Job seekers who have physical disability and need long-term support, i.e. program=yes, benefit most. As a low number of working hours is a major barrier for these people, improving work capacity would increase their employability.

5.5 Answers to domain questions

1) Our recommendations would increase employability. Figure 4 show that AUUC of MCT is positive. If job seekers followed

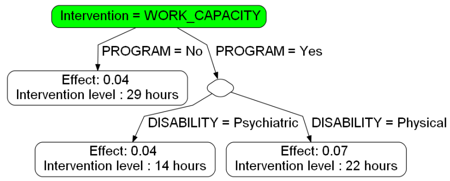

Figure 9: Causal tree for work capacity intervention.

our recommendations to upgrade the most effective factor, their employment chance would increase. The expected employability increase is between 8.5% to 16.5% as showed in Figure 5.

2) Factors for intervention. Motivation and work capacity are two causal factors of employment status with positive effects. Their intervention would increase employment chance. Intervention on education is another option but this factor does not work for immature and old job seekers. The most effective factor varies among job seekers as it depends on their characteristics.

3) Best intervention levels. Figures 7, 8, 9 show that recommended intervention levels to achieve maximum causal effects are not the highest levels. This implies non-linear relationships between intervention amount and employability increase. Upgrading a factor beyond recommended levels is not necessary. The best intervention levels also vary among heterogeneous job seekers.

4) Degree of employability increase Causal effects from causal trees allow us to determine how much employability of a job seeker would increase if he/she adopted our recommendation. The larger causal effect, the higher employability increase.

5.6 Practical Implications

1) It is unreasonable to intervene any factors. Not all factors can be intervened to increase employability of disabled job seekers. Three causal factors: motivation, education, and work capacity, can help us achieve this goal. Intervention on other factors would not affect employment status, thus wasting efforts of job seekers.

2) Two ways of selecting factors for intervention. Job seekers can select the most effective factor for intervention to improve their employment chance. Alternatively, they can select a factor whose recommended intervention level is achievable to them.

3) Personalized intervention recommendation. The most effective intervention, i.e. a factor and its best intervention level, to achieve maximum employability increase varies among job seekers. The best level may not be the highest level. Recommended levels can be used to design upskilling programs tailored to individual job seekers for desired employability increase.

4) Adverse effects of education. Upgrading education generally helps to improve employment chance. For old job seekers, however, it is advised that they should not improve this factor as its intervention would reduce their employability.

5) Resource allocation for employment assistance. Causal effect indicates how much employment chance would increase if job seekers adopted recommendations. Providers can proceed to upskill those with high causal effects. For people with low causal effects, we need to explore other factors that are not included in our analysis. Then our method is used to determine the most effective factors for increasing employability of these people, and DES providers can assist them in upgrading recommended factors.

6) Actionable recommendations. Causal factors and causal trees generated by our method are interpretable. This makes our recommended interventions actionable in the DES context where interventions require efforts of both providers and job seekers.

6 RELATED WORK

Disability employment. A growing literature focuses on employment barriers for disabled job seekers. Various factors are found to influence employment chance of job seeker [1, 16, 21]. Association-based methods are used to detect correlations in data collected from studies designed for research. Our work supplements this literature by developing a causality-based method for recommending the most effective interventions using observational data. Recently, a method for recommending interventions for disability employment is proposed [33], but it only works with binary interventions.

Employment recommendation. Popular recommendation systems deployed for the employment service are candidate recommendation [2, 3, 26], or job recommendation [3, 11, 12]. Both collaborative filtering (CF) and content-based (CB) methods are used to develop these systems. Hybrid approaches are also applied to leverage strengths of CF and CB[22, 34]. Skill recommendation is another related area where relevant skills are recommended for a job or a career path [7, 9, 31]. The target problems of these recommendation systems do not involve counterfactual reasoning. Their goal is to optimize employment status based on association between factors and the outcome [33]. Our problem focuses on *increase* in employment chance and needs to address counterfactual questions.

Treatment effect heterogeneity (TEH) & uplift modelling. TEH refers to the fact that a treatment has different effects on different individuals[4]. In precision medicine, TEH allows doctors to prescribe a treatment to patients that most benefit from it. In business analytics, TEH, also known as uplift modelling, is used to estimate effects of a promotion on a business outcome.

There are two groups of methods for estimating TEH. The first group works with randomized trials [25, 29, 30, 36], which is not applicable to our problem since our data are observational. The second group is developed to deal with observational data with tree based [4, 5, 28, 29, 33, 37] and non tree-based methods [18, 32, 35]. The former is more related to our work since our proposed method is also tree-based to achieve interpretability. However, most of them only work with binary treatments. To recommend the best intervention levels, these methods need to binarize causal factors in all possible ways and build tree models for each binarization. This approach has two disadvantages: 1) high computational cost 2) low interpretability since we need to build many trees. Those disadvantages become more prohibitive with continuous factors. We need to binarize the factors at all values to build tree models for recommending the best intervention levels for those factors.

Some tree-based methods are proposed to estimate treatment effect heterogeneity for multivalued treatments, but they are still only applied to randomized trials [39, 40]. While most current methods assume known causal factors, our problem requires causal factors to be detected from observational data.

7 CONCLUSION

We present the first study of recommending the most effective factor with its best intervention level to increase employment chance of disabled people. We develop a causality-based method with two stages where a counterfactual framework is combined with a machine learning approach to address counterfactual questions. Model performance is evaluated with both synthetic datasets and a case study. Results show consistent promising performance of improving employment chance and discovering the heterogeneity of job seekers. We disclose causal factors that disabled job seekers should intervene to increase their employability. Causal effects for individuals, which are estimated by interpretable causal trees, vary among job seekers. We also answer domain questions from our industry partner and provide practical implications for DES.

Limitations. First, we assume causal effect of a factor is independent among individuals. In certain conditions, increase in employment chance of one person may affect employment chance of others. Second, interactions among multiple causal factors are not considered in our work. Job seekers may intervene multiple factors at the same time and the combined effect can be different from the sum of separate effects. Finally, our case study is based on data collected from Disability Employment Services in one country. Our analysis may be generalized to other settings with caution.

REFERENCES

[1] TJ Achterberg, H Wind, AGEM De Boer, and MHW Frings-Dresen. 2009. Factors that promote or hinder young disabled people in work participation: a systematic review. *Journal of occupational rehabilitation* 19, 2 (2009), 129–141.
[2] Shaha T Al-Otaibi and Mourad Ykhlef. 2012. A survey of job recommender systems. *International Journal of Physical Sciences* 7, 29 (2012), 5127–5142.
[3] Nikolaos D Almalis, George A Tsihrintzis, Nikolaos Karagiannis, and Aggeliki D Strati. 2015. FoDRA—A new content-based job recommendation algorithm for job seeking and recruiting. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 1–7.
[4] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.
[5] Susan Athey, Guido W Imbens, et al. 2015. *Machine learning for estimating heterogeneous causal effects*. Technical Report.
[6] Susan Athey, Julie Tibshirani, Stefan Wager, et al. 2016. Solving heterogeneous estimating equations with gradient forests. *arXiv preprint arXiv:1610.01271* (2016), 168–189.
[7] Azuraini Abu Bakar and Choo-Yee Ting. 2011. Soft skills recommendation systems for IT jobs A Bayesian network approach. In *2011 3rd Conference on Data Mining and Optimization (DMO)*. IEEE, 82–87.
[8] Steen Bengtsson and Nabanita Datta Gupta. 2017. Identifying the effects of education on the ability to cope with a disability among individuals with disabilities. *PLoS One* 12, 3 (2017), e0173659.
[9] Vachik S Dave, Baichuan Zhang, Mohammad Al Hasan, Khalifeh AlJadda, and Mohammed Korayem. 2018. A combined representation learning approach for better job and skill recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1997–2005.
[10] Alex A Freitas. 2004. A critical review of multi-objective optimization in data mining: a position paper. *ACM SIGKDD Explorations Newsletter* 6, 2 (2004), 77–86.
[11] Shiqiang Guo, Folami Alamudun, and Tracy Hammond. 2016. RésuMatcher: A personalized résumé-job matching system. *Expert Systems with Applications* 60 (2016), 169–182.
[12] Anika Gupta and Deepak Garg. 2014. Applying data mining techniques in job recommender system for considering candidate job preferences. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 1458–1465.
[13] Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*. 1–13.
[14] Behram Hansotia and Brad Rukstales. 2002. Incremental value modeling. *Journal of Interactive Marketing* 16, 3 (2002), 35.
[15] Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 4 (2003), 1161–1189.
[16] Anja Holwerda, Jac JL van der Klink, Michiel R de Boer, Johan W Groothoff, and Sandra Brouwer. 2013. Predictors of work participation of young adults with mild intellectual disabilities. *Research in developmental disabilities* 34, 6 (2013), 1982–1990.
[17] Marlene Krasovitsky, Angela Zaballa, Do Purcell, A Mitchell, T Davidson, C Gebicki, L Drakopoulos, L Gunn, M Wadhera, W La, et al. 2016. Willing to work: National inquiry into employment discrimination against older Australians and Australians with disability.
[18] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.
[19] Marco Laumanns, Lothar Thiele, Kalyanmoy Deb, and Eckart Zitzler. 2002. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary computation* 10, 3 (2002), 263–282.
[20] Jiuyong Li, Lin Liu, and Thuc Duy Le. 2015. *Practical Approaches to Causal Relationship Exploration*. Springer International Publishing, Cham, Cham. https://doi.org/10.1007/978-3-319-14433-7
[21] Sally Lindsay. 2011. Discrimination and other barriers to employment for teens and young adults with disabilities. *Disability and rehabilitation* 33, 15-16 (2011), 1340–1350.
[22] Yao Lu, Sandy El Helou, and Denis Gillet. 2013. A recommender system for job seeking and recruiting website. In *Proceedings of the 22nd International Conference on World Wide Web*. 963–966.
[23] Judea Pearl. 2009. *Causality: Models, reasoning, and inference, second edition*. 1–464 pages. https://doi.org/10.1017/CBO9780511803161
[24] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.
[25] Piotr Rzepakowski and Szymon Jaroszewicz. 2010. Decision trees for uplift modeling. In *2010 IEEE International Conference on Data Mining*. IEEE, 441–450.
[26] Walid Shalaby, BahaaEddin AlAila, Mohammed Korayem, Layla Pournajaf, Khalifeh AlJadda, Shannon Quinn, and Wlodek Zadrozny. 2017. Help me find a job: A graph-based approach for job recommendation at scale. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1544–1553.
[27] Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. MIT press.
[28] Falco J Bargagli Stoffi and Giorgio Gnecco. 2020. Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms. *International Journal of Data Science and Analytics* 9, 3 (2020), 315–337.
[29] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10, 2 (2009).
[30] Christopher Tran and Elena Zheleva. 2019. Learning triggers for heterogeneous treatment effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5183–5190.
[31] Kush R Varshney, Jun Wang, Aleksandra Mojsilović, Dongping Fang, and John H Bauer. 2013. Predicting and recommending skills in the social enterprise. In *Seventh International AAAI Conference on Weblogs and Social Media*.
[32] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
[33] H. Xuan Tran, T. Duy Le, J. Li, L. Liu, J. Liu, Y. Zhao, and T. Waters. 2020. Intervention Recommendation for Improving Disability Employment. In *2020 IEEE International Conference on Big Data (Big Data)*. 1671–1680. https://doi.org/10.1109/BigData50022.2020.9378350
[34] Shuo Yang, Mohammed Korayem, Khalifeh AlJadda, Trey Grainger, and Sriraam Natarajan. 2017. Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach. *Knowledge-Based Systems* 136 (2017), 37–45.
[35] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.
[36] Łukasz Zaniewicz and Szymon Jaroszewicz. 2013. Support vector machines for uplift modeling. In *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 131–138.
[37] Achim Zeileis, Torsten Hothorn, and Kurt H. 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17, 2 (2008), 492–514.
[38] Weijia Zhang, Thuc Duy Le, Lin Liu, and Jiuyong Li. 2018. Estimating heterogeneous treatment effect by balancing heterogeneity and fitness. *BMC bioinformatics* 19, 19 (2018), 61–72.
[39] Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. A practically competitive and provably consistent algorithm for uplift modeling. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1171–1176.
[40] Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 588–596.

8 SUPPLEMENT

8.1 Model construction

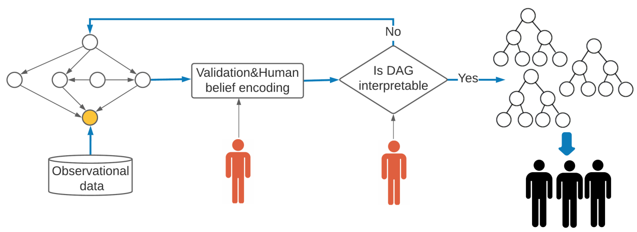

Figure 10: Main steps of building a MCT model.

8.1.1 Building MCT model. The process of building an MCT model includes three main steps. The first step is causal discovery where a causal DAG is inferred from disability employment data. Causal relationships in the DAG are validated by domain experts in the second step, i.e. validation and human belief encoding. If some relationships are unreasonable, we fix them in the process of learning causal structures. In the third step, a causal tree is constructed for each causal factor of employment status. These trees are used to recommend the most effective interventions to job seekers.

8.1.2 Estimating maximum causal effect and the best intervention level. Algorithm 3 presents main steps to estimate a maximum causal effect and the best intervention level of causal factor F for subgroup $\mathcal{G}$.

Algorithm 3 Estimate maximum causal effect τ^* and the best intervention level θ^* of factor F for subgroup $\mathcal{G}$.

Input: Subgroup $\mathcal{G}$, causal factor F.
Output: $\widehat{\tau^*}$ and $\widehat{\theta^*}$.

```
1: Function EstimateMaxEffect(G, F)
2: for each θ ∈ G(F) do
3:     estimate causal effect τ̂_G(θ) by Equation 5
4: end for
5: return τ̂* = max τ̂_G(θ) and θ̂* = argmax_θ(τ̂_G(θ)).
6: End
```

Approximate binning strategy. Normally, we need to scan every value of continuous factor F to find the best intervention level and propensity scores have to be re-estimated for each test. To reduce the computational cost, we employ an approximate binning strategy to reduce the number of candidates to scan through. Data points are sorted by F in ascending order and divided into unit blocks whose distribution of outcome Y is assumed to be uniform. Two consecutive blocks are merged together if they have similar distribution of Y. T-test is used to compare distributions. This procedure continues until no merging is performed. After the merging, possible best intervention levels are the values at block borders. These values are used globally for all splittings. Propensity scores of these intervention levels are estimated and re-used during the tree construction process. This strategy is applied to continuous causal factors before building causal trees for those factors.

8.1.3 Causal conditions for DAG construction. Below are additional conditions for building a causal DAG from data.

Definition 5. Faithfulness condition. Causal graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ and joint distribution P($\mathbf{V}$) satisfy the faithfulness condition iff all conditional independence relationships in P($\mathbf{V}$) are entailed by the Markov condition. This assumption implies that if two variables are probabilistically dependent, there must be an edge connecting these variables in the $\mathbf{G}$.

Definition 6. Causal sufficiency. Given a pair of variables observed in a dataset, their common causes are also observed.

8.2 Generating synthetic data

2000 observations are generated with 20-50 attributes $\mathbf{X}$ describing job seekers, causal factor F for intervention, and employment status Y. Attributes $\mathbf{X}$ follow normal distribution, causal factor F follows truncated normal distributions or Poisson distribution, and employment status Y follows Bernoulli distribution. Distribution parameters for F and Y depend non-linearly on attributes $\mathbf{X}$ and interactions among the attributes. Y also depends on F and the interaction of this factor with attributes $\mathbf{X}$. The data generation process simulates real conditions in the DES sector that upgrading factor F would increase the chance to get employed. Table 6 presents four scenarios of generating data with increased complexity levels.

To be used for baseline methods, factor F needs to be binarized. We select 50 breakpoints that divide F into equal bins. These breakpoints are used to binarize factor F to create 50 derived binary datasets from an original dataset. A derived binary factor F has value of 0 if its original value is smaller than a breakpoint. Otherwise, the binary factor has a value of 1.

8.3 Data split and model training

Synthetic datasets are randomly split into 70% for training models and 30% for evaluating performance with five splits. To determine hyperparameter $\boldsymbol{\epsilon} = (\epsilon_{ef}, \epsilon_{ed}, \epsilon_{ld})$, we conduct grid-search where $\epsilon_{ef} \in [0, 0.15]$, $\epsilon_{ed} \in [0, 0.5]$, $\epsilon_{ld} \in [0, 0.5]$ with the interval of 0.05. We limit the maximum value of ϵ_{ef} to 0.15 since we need a more strict criterion on effect fitness for selecting an optimal split. The metric for hyperparameter selection is $H = \frac{AUUC}{Max(AUUC)} + \frac{Kendall}{Max(Kendall)}$, where $AUUC$ is Area Under Uplift Curve, $Max(AUUC)$ is the maximum value of $AUUC$ among all results, $Kendall$ is Kendall correlation coefficient, $Max(Kendall)$ is the maximum value of $Kendall$ among all results. The metric is to select a hyperparameter of a model that achieves the highest combination of $AUUC$, i.e. employment increase, and $Kendall$, i.e. heterogeneity discovery. Experiments on the real dataset are also conducted in the same way.

8.4 Implementation

Our method is implemented using Python. The codes are available at *https://github.com/trxuanha/maximumcausaltree.* We use the PC algorithm in package pcalg for causal discovery.

To compare with baseline methods, we use the existing implementation of the baseline methods for evaluation. The code is located at *https://github.com/susanathey/causalTree.*

Table 6: Scenarios of generating synthetic data

Attributes $\mathbf{X}$	Causal factor F	Employment status Y	Comments
20 variables $\sim Normal(0,1)$	$Max(\boldsymbol{\alpha}(\boldsymbol{\beta}+0.35)\mathbf{X}^2, 0)$	$\sim Bernoulli(\frac{1}{1+e^{-(\boldsymbol{\gamma}\boldsymbol{\delta}\mathbf{X}+9F)^2+2}})$	$\alpha, \gamma, \eta \sim Bernoulli(p)$ $\beta, \delta, \theta \sim Normal(0,1)$
	$Max(\boldsymbol{\alpha}(\boldsymbol{\beta}+0.35)\mathbf{X}^2$ $+tan(\sum_{k,p}^{20}\eta(\theta+0.1)X_kX_p), 0)$	$\sim Bernoulli(\frac{1}{1+e^{-(\boldsymbol{\gamma}\boldsymbol{\delta}\mathbf{X}+9F)^2-tan(\sum_{k,p}^{20}\eta\theta X_kX_pF)+3.5}})$	
50 variables $\sim Normal(0,1)$	$Max(\boldsymbol{\alpha}(\boldsymbol{\beta}+0.35)\mathbf{X}^2$ $+tan(\sum_{k,p}^{50}\eta(\theta+0.1)X_kX_p), 0)$	$P=\frac{1}{1+e^{-(\boldsymbol{\gamma}\boldsymbol{\delta}\mathbf{X}+9F)^2-5.5tan(\sum_{k,p}^{50}\eta\theta X_kX_pF)+2}}$ $Y_i \sim Bernoulli(P\mathbb{I}_{\{P<=0.7\}}+0.7\mathbb{I}_{\{P>0.7\}})$	
50 variables $\sim Normal(0,1)$	$\sim Poisson(Max(\boldsymbol{\alpha}(\boldsymbol{\beta}+0.35)\mathbf{X}^2$ $+tan(\sum_{k,p}^{50}\eta(\theta+0.1)X_kX_p), 0))$	$P=\frac{1}{1+e^{-(\boldsymbol{\gamma}\boldsymbol{\delta}\mathbf{X}+9F)^2-5.5tan(\sum_{k,p}^{50}\eta\theta X_kX_pF)+2}}$ $Y_i \sim Bernoulli(P\mathbb{I}_{\{P<=0.7\}}+0.7\mathbb{I}_{\{P>0.7\}})$	

Note: $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\delta}$ has the same dimension as $\mathbf{X}$.
The dimension of $\boldsymbol{\eta}, \boldsymbol{\theta}$ equals the total number of degree-2 polynomial combinations among attributes of $\mathbf{X}$.
For example if $\mathbf{X} = \{X1, X2, X3\}$, the total number of degree-2 polynomial combinations among attributes of $\mathbf{X}$ is 6.
This is because we have 6 combinations: $X1, X2, X3, X1X2, X1X3, X2X3$.
$\alpha \in \boldsymbol{\alpha}, \beta \in \boldsymbol{\beta}, \gamma \in \boldsymbol{\gamma}, \delta \in \boldsymbol{\delta}, \theta \in \boldsymbol{\theta}, \eta \in \boldsymbol{\eta}$.