Pre-submitted to arXiv

An interpretable machine learning framework for modelling human decision behavior

Mengzhuo Guo
Xi'an Jiaotong University, City University of Hong Kong, mengzhguo2-c@my.cityu.edu.hk,

Qingpeng Zhang*
City University of Hong Kong, qingpeng.zhang@cityu.edu.hk,

Xiuwu Liao
Xi'an Jiaotong University, liaoxiuwu@mail.xjtu.edu.cn,

Youhua Chen
City University of Hong Kong, youhchen@cityu.edu.hk,

Machine learning has recently been widely adopted to address the managerial decision making problems. However, there is a trade-off between performance and interpretability. Full complexity models (such as neural network-based models) are non-traceable black-box, whereas classic interpretable models (such as logistic regression) are usually simplified with lower accuracy. This trade-off limits the application of state-of-the-art machine learning models in management problems, which requires high prediction performance, as well as the understanding of individual attributes' contributions to the model outcome. Multiple criteria decision aiding (MCDA) is a family of interpretable approaches to depicting the rationale of human decision behavior. It is also limited by strong assumptions (e.g. preference independence). In this paper, we propose an interpretable machine learning approach, namely **N**eural **N**etwork-based **M**ultiple **C**riteria **D**ecision **A**iding (NN-MCDA), which combines an additive MCDA model and a fully-connected multilayer perceptron (MLP) to achieve good performance while preserving a certain degree of interpretability. NN-MCDA has a linear component (in an additive form of a set of polynomial functions) to capture the detailed relationship between individual attributes and the prediction, and a nonlinear component (in a standard MLP form) to capture the high-order interactions between attributes and their complex nonlinear transformations. We demonstrate the effectiveness of NN-MCDA with extensive simulation studies and two real-world datasets. To the best of our knowledge, this research is the first to enhance the interpretability of machine learning models with MCDA techniques. The proposed framework also sheds light on how to use machine learning techniques to free MCDA from strong assumptions.

Key words: Explainable artificial intelligence; Interpretable machine learning; Interpretability; Data-driven decision making; Intelligent model; Decision analysis; Mutiple criteria decision analysis.

* Corresponding author

1. Introduction.

Machine learning has recently been widely adopted to address challenging decision making problems in a variety of managerial contexts like marketing (Cui and Curry 2005, Cui et al. 2006), credit-risk evaluation (Baesens et al. 2003) and healthcare management (Gartner et al. 2015). Many machine learning models, such as support vector machines (SVMs) (Cortes and Vapnik 1995), boosted trees (Friedman 2001) and neural network-based methods (Rumelhart et al. 1988, LeCun et al. 2015), are applied to diverse real-world prediction problems due to their capacity to analyze high-dimensional data. Although higher complexity usually brings higher accuracy, it comes at the expense of interpretability (Lou et al. 2012). In practice, model interpretability is as important as (if not more important than) accuracy in many mission-critical applications such as clinical decision-making, in which the understanding of how the model makes the prediction is the key to facilitate physicians to trust the model and utilize the prediction results (Caruana et al. 2015, Fox et al. 2007). Recently, technology giants like Google, IBM, and Microsoft, have been investigating on the techniques in enhancing the model interpretability (Mohseni et al. 2018). As stated in a comprehensive overview conducted by Mr. David Gunning, the program manager in the Information Innovation Office (I2O) of the Defense Advanced Research Projects Agency (DARPA), "*machine learning models are opaque, non-intuitive and difficult for people to understand*" (Gunning 2017). DARPA has since funded for developing interpretable machine learning techniques among academics. In the latest budget plan of DARPA, explainable artificial intelligence (XAI) has been listed as the key funding area in the fiscal year 2019-2020, with the total amount of 26.05 million US dollars [1].

1.1. Benefits of interpretable models

Both machine learning and management research could benefit from the interpretability of models. First, an interpretable model is trustworthy because it exploits some patterns and rules that are consistent with prior human knowledge and experiences. The unreasonable learned patterns and rules can be easily identified and corrected by a decision maker (DM). If a model tends to make mistakes that can be easily to be classified accurately by the DM, it would require his/her supervisions of modification (Ribeiro et al. 2016). Second, an interpretable model helps in understanding causality (Miller 2018). Interpretable models

[1] `https://www.darpa.mil/about-us/budget`

can extract the associations between predictors and predictions, which can facilitate the downstream managerial decision making. Third, an interpretable model incorporates the DM's domain knowledge. A DM usually possesses rich domain knowledge but not technical skills to construct a model. An interpretable model can be used to learn a DM's decision behavior through tuning key parameters and then provide in-depth understanding of the data and patterns (Aggarwal and Fallah Tehrani 2019).

1.2. Multiple criteria decision aiding

Multiple criteria decision aiding (MCDA)[2] has been a fast growing area of operational research during the last several decades (Dyer et al. 1992, Figueira et al. 2005, Wallenius et al. 2008, Saaty 2013, Ramesh et al. 1988, 1989). It involves a finite set of alternatives (e.g. actions, items, policies) that are evaluated from a set of conflicting multiple criteria or attributes[3]. The DM's decision is driven by his/her underlying *global value (utility) function* (Keeney 1976, Keeney and Raiffa 1993). This global value measures the DM's desirability for an alternative and can be disaggregated into a set of per-attribute *marginal value functions* that represent the DM's evaluation of the corresponding attribute (Kadziński et al. 2017). These marginal value functions can be learned by the DM's judgments on learning examples (e.g. pairwise comparison between two alternatives). Once the marginal value functions are deciphered, we can understand the decision making rationale, based on which we can predict the judgment of the DM. This process is referred as the preference disaggregation approaches of MCDA.

Many machine learning framework can help MCDA accomplish the learning objectives because both of them aim to learn a decision model from data. Thus, MCDA and machine learning naturally have reciprocal interactions (Doumpos and Zopounidis 2011). MCDA and machine learning are integrated in two directions. First, we can apply machine learning techniques to various tasks in a decision aiding context, such as learning to rank, multi-label classification, etc. The opposite direction is to implement MCDA concepts in a machine learning framework. It is a tendency that utilize MCDA approaches to adapt the machine learning models to various topics, such as feature selection and extraction, pruning decision

[2] Multiple criteria decision aiding is also named as multiple criteria decision making. In this paper, we use multiple criteria decision aiding (the "European school") for consistency (Vincke 1986).

[3] In machine learning, criteria refer to attributes or features with preference order scales (Corrente et al. 2013). For consistency, we use "attribute" in this paper.

rules and multiple objective optimization. Our work belongs to the second stream. We aim to construct a hybrid model, which utilizes value function-based preference disaggregation approaches of MCDA to enhance the interpretability of "black-box" machine learning models.

The motivation of introducing the value function-based preference disaggregation approaches of MCDA to machine learning stems from its powerful capacity in depicting the human decision-making process. The deciphered marginal value functions reveal the rationale of DM's judgment, and thus provide convincing evidence to assist comprehending the decision making behavior (Aggarwal and Fallah Tehrani 2019, Lou et al. 2012). Our task of learning an interpretable model is essentially to capture the characteristics of the marginal value functions, based on which we obtain a certain degree of interpretability. This study is different from the statistical models for management problems, in that we utilize the characteristics of the marginal value function (instead of a single coefficient) to represent the effect of each attribute on the outcomes. For example, suppose that a hypothetical DM's preference system is composed of four marginal value functions in Figure 1. We can analyze the DM's preference from the following perspectives. First, we focus on the ranges of the marginal values. If the marginal values are close to 0 (see the marginal value function 1), it indicates that the corresponding attribute is not important to the DM or we have wrongly captured its characteristic. Further interaction with the DM is needed to determine whether we keep this variable or calibrate the model. Different from the statistical model selection methods (e.g.: LASSO, Bayesian information criteria (Friedman et al. 2001)), through incorporating the DM's domain knowledge, an interactive model calibration process is invoked (Stewart 1993, Wallenius et al. 2008, Doumpos and Zopounidis 2011). Second, the increasing and decreasing tendencies of the marginal value function curves unveil the change of the DM's preference. Moreover, the negative and positive marginal values directly show the negative and positive effects of the attributes on the outcomes (see the marginal value functions 2 and 3). Statistical models usually generate a fixed coefficient that cannot capture such preference inflexion points. Third, the convexity and concavity of the marginal value function are crucial for interpreting the DM's rational behavior in decision-making process (see the marginal value function 4). To summarize, different from traditional statistical models, MCDA aims to extract more interpretable patterns of the DM's behavior and builds a solid link between the underlying model and

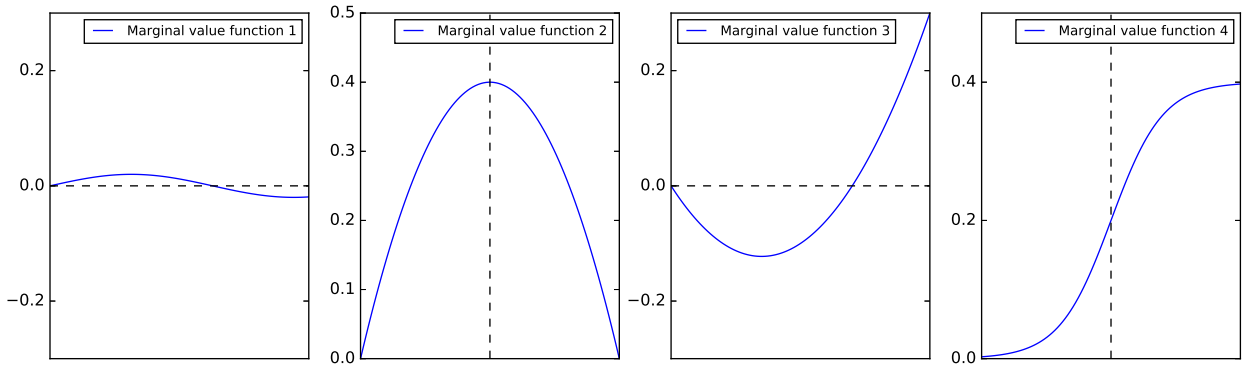

Figure 1 **Exemplary marginal value funcions.**

the actual decision making processes, and thus facilitates the effective use of the model in real-world decision makings.

1.3. An overview of this paper.

This paper proposes a framework for a **N**eural **N**etwork-based **M**ultiple **C**riteria **D**ecision **A**iding (NN-MCDA) approach. NN-MCDA combines an additive model and a fully-connected multilayer perceptron (MLP) to achieve both model interpretability and complexity. The additive model is learned from the value function-based preference disaggregation models of MCDA. It uses marginal value functions to approximate the relationship between the outcome and individual attributes whereas the MLP is used to capture the high-order correlations between attributes in the model. We estimate the parameters in the model under a neural network framework that automatically balances the trade-off between two components.

We tested our proposed model using a set of synthetic datasets and two real datasets. Specifically, the simulation experiments respectively show the impact of pre-defined parameters on the model and the goodness of the model when data is either extremely complex or simple. Two real datasets on ranking universities regarding employment reputation and predicting the risk for geriatric depression are utilized to illustrate the proposed model in real cases. We explain the obtained models and compare them to other interpretable models, i.e., GAM and logistic regression models.

The contributions of this paper are fourfold. First, we advocate a new perspective of an interpretable model that both quantifies the impact of individual attributes on the outcome and captures the possible high-order correlations between attributes in the model. It helps

the DM to understand the main effect of single attribute and to make better decisions. Second, to the best of our knowledge, this paper is the first pilot work that introduces the value function-based preference disaggregation approaches of MCDA to the machine learning models to enhance the model interpretability. The trained parameters in the proposed framework determine the shape for marginal value functions in the additive models. The proposed model is free from preference independence, preference monotonicity, and small learning set assumptions in MCDA approaches, which makes MCDA approaches more general and practical for real-world management problems. Third, we examine the model effectiveness given different model parameters and datasets. The empirical conclusions about the relationships between model interpretability and data complexity are managerially intuitive for the future researches. Forth, the proposed framework is flexible and extendible, especially the nonlinear part, which can be modified or replaced by other models according to different types of data. Our work is intuitive for developing interpretable models for both management and computational science.

The rest of the paper is organized as follows. We discuss the related work in Section 2. In Section 3, we introduce the framework for the proposed interpretable model. The simulation and real case experiments are presented in Section 4 and some discussions about the proposed framework is provided in Section 5. We conclude the paper in Section 6.

2. Related work.

2.1. Value function-based preference disaggregation approach of MCDA.

The value function-based preference disaggregation approaches of MCDA provide explicit marginal value functions and numerical scores. A DM can understand the importance of a particular attribute and how the individual attributes contribute to the final decision. This procedure encourages the DM to participate in the decision making process and it provides a comprehensive preference model (Corrente et al. 2013). These approaches have been successfully applied to many scenarios, such as consumer preference analysis (Hauser 1978), financial decisions (Zopounidis et al. 2015), nano-particles synthesis assessment (Kadziński et al. 2018) and territorial transformation management (Ciomek et al. 2018). However, the applications of value function-based preference disaggregation approaches are limited due to some strong assumptions, such as (1) preference independence, (2) monotonic preference, and (3) small set of alternatives.

Recently, many novel models haven been proposed to generalize the value function-based preference disaggregation approaches of MCDA. Preference independence allows the model to be additive. Considering interacted attributes, Angilella et al. (2010) utilize a fuzzy measure to model the preference system where the alternatives are now evaluated in terms of the Choquet integral. However, it is difficult for the DM to understand the impact of individual attribute evaluated from the Choquet integral. Angilella et al. (2014) account for positive and negative interactions among attributes, and add an interaction term to the additive global value function for each alternative. They require the DM to provide some knowledge about the interacted pairs that are mined by the models. These studies only consider the interaction between pairs of attributes because higher-order interactions require more cognitive efforts and more computational cost.

The majority of existing researches assume the marginal value functions are monotonic piece-wise linear. This assumption reduces the model complexity, but it fails to describe preference inflexions. Addressing this problem, Ghaderi et al. (2017) and Liu et al. (2019) relax this assumption and constrain on variations of the slope to obtain non-monotonic marginal value functions without serious over-fitting problem. Both of their approaches obtain non-smooth value functions which are difficult to interpret attitudes towards risks due to the use of non-derivative functions. Since a differetiable margianl value function is essential to analyze consumer behavior, Sobrie et al. (2018) utilize semidefinite programming to infer the key parameters for polynomial marginal value functions. It gives a more flexible and interpretable preference model. However, it still assumes that the DM preference is monotonic.

The monotonic piece-wise form of the marginal value functions has a low expressibility for large learning sets (Sobrie et al. 2018). Nowadays, MCDA approaches are expected to deal with large amount of data in many disciplines (Pelissari et al. 2019, Liu et al. 2019). Liu et al. (2019) embed the MCDA approach into a regularization framework to approximate marginal value functions in any piece-wise linear shapes, and provide efficient algorithms to handle larger learning sets.

Most existing researches focus on expanding the MCDA approaches from only one perspective. Comparing with these recent advances, the proposed framework tries to solve all aforementioned limitations of MCDA by providing a non-monotonic, smoother, and more powerful MCDA approach for real-world applications considering more complex decision making scenarios.

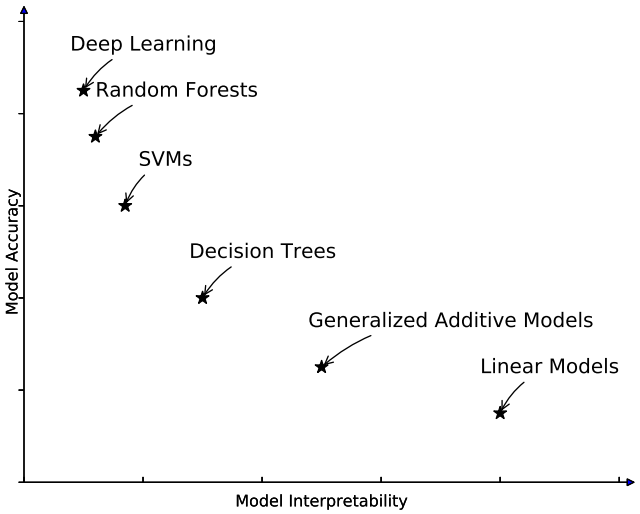

Figure 2 Prediction accuracy vs. interpretability for some learning techniques.

Note. This figure is based on a DARPA report conducted by Gunning (2017). Since the proposed model imports neural network, we believe it can theoretically achieve a result that are at least as good as random forests. Moreover, the proposed model uses a linear additive model, it also has a good interpretability.

2.2. Interpretable models.

There is usually a trade-off between model interpretability and prediction accuracy (shown in Figure 2). Interpretable machine learning, or XAI, aims to create a suite of techniques that produce more explainable models while maintaining a high accuracy (Gunning 2017).

Generalized additive model (GAM) uses a link function to build a connection between the mean of the prediction and a smooth function of the predictors (Hastie and Tibshirani 1986). It is good at both dealing with and presenting the nonlinear and non-monotonic relationship between the predictors and the prediction (Lou et al. 2012). Therefore, GAM is usually more accurate than linear additive models. Although GAM does not outperform full complexity models, it possesses more interpretability than these "black-box" models. Lou et al. (2013) explore the co-effect of pairwise interactions and apply the improved GAM to predicting pneumonia risk and 30-day readmission. This model helps the DM (physician) to find useful patterns in the data and quantifies the contributions of individual attributes. Based on these promising results, they argue that it is necessary to develop more interpretable models in mission-critical applications such as management problems (Caruana et al. 2015).

Another solution is to infer a new model to approximate the true black-box model. The new model may not be as accurate as the original black-box model, but can iden-

tify patterns and rules to explain how the predictions are made. In Baesens et al. (2003), explanatory rules are extracted to help the credit-risk managers in explaining their decisions. Similarly, Letham et al. (2015) discretize a high-dimensional attribute space into a series of simpler interpretable *if-then* statements. They firstly make predictions using complex machine learning techniques and then use Bayesian rule lists to reconstruct the predictions. Given approximately accurate predictions, the obtained model is more interpretable.

According to Ribeiro et al. (2016), why and how the model produces that prediction are important for the DM to trust the underlying model. An interpretable model should enable to answer these questions and give the reasons behind a prediction. In this regard, they develop an algorithm named LIME which approximates a prediction locally with a simper model, for instance a linear model that is easier to interpret. It is extensible to explain the predictions of any model in an interpretable manner.

2.3. Machine learning in MCDA.

There have been a few attempts to integrate the machine learning algorithms with MCDA. In a pioneering work by Wang and Malakooti (1992), a single-layered feed-forward artificial neural network is proposed to learn MCDA objectives. The advantages of neural networks are that they are independent of functional forms. However, it only gives a final recommendation without any interpretable marginal value functions or patterns. Doumpos and Zopounidis (2011) explore the differences and similarities between machine learning and MCDA. Although there are several studies introducing MCDA into machine learning models, few utilize the MCDA concepts to enhance machine learning models' interpretability.

As a new sub-field of machine learning, preference learning has attracted extraordinary attention from the MCDA community. Corrente et al. (2013) explore the relationship between MCDA and preference learning. They find that the higher performance of machine learning models is usually associated with lower degree of interpretability, which negatively affects the confidence in the employment of machine learning models in scenarios where we need to understand the underlying process. A latest study utilizes preference learning to model human decision behavior under a MCDA framework. Such a model can facilitate the understanding of the DM's behavior by tuning well-defined model parameters (Aggarwal and Fallah Tehrani 2019).

3. Framework for the intelligible model.

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_1^N$ be the training dataset of size N, $\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,n})^T$ be the i-th attribute vector with n attributes[4], and y_i be the target/response value. In this study, we consider a binary classification problem where $y_i \in \{0, 1\}$. The proposed framework can be easily extended to multi-classification and regression problems.

3.1. The additive model.

The value function-based preference disaggregation approaches of MCDA assume that for each attribute vector $\mathbf{x}_i \subseteq \mathbb{R}^n$, there is a global value function in the following form:

$$F(\mathbf{x}_i) = w_1 \cdot v_1(x_{i,1}) + w_2 \cdot v_2(x_{i,2}) + \cdots + w_n \cdot v_n(x_{i,n}) = \sum\nolimits_{j=1}^{n} w_j v_j(x_{i,j}) \tag{1}$$

where $w_j \geq 0, j = \{1, 2, \cdots, n\}$ represents the importance of the j-th attribute and $v_j(\cdot)$ is a marginal value function. Note that we reply the shape and positive/negative effect of the marginal value function to capture the contribution of individual attributes. Thus we set the weight w_j to be positive to represent the relative importance of the j-th attribute, which can positively or negatively affect the global value. The *global value function* $F(\cdot)$ linearly sums contributions of individual attributes[5].

Although the global value function is in an additive linear form, the marginal value functions themselves can be in any forms, often nonlinear. It has been recognized that the preference in human decision behaviors is **rational**, and thus the marginal value functions should be stable and smooth. In the literature, the marginal value function can be in a simple linear (weighted sum) form (Saaty and Decision 1990, Saaty 2013, Korhonen et al. 2012), monotonic and non-monotonic piecewise linear forms (Stewart 1993, Jacquet-Lagreze and Siskos 2001, Greco et al. 2008, Ghaderi et al. 2017, Liu et al. 2019), and monotonic polynomial form (Sobrie et al. 2018). To capture the first-order (e.g. monotonicity) and second-order (e.g. marginal rate in substitution) derivative patterns of the attributes' contributions to the prediction, we extend and generalize state-of-the-art MCDA models (Liu et al. 2019, Sobrie et al. 2018) to allow the marginal value function in any polynomial forms. In this paper, we allow the j-th marginal value function to be in a smooth and non-monotonic form of D_j degrees:

$$v_j(x_{i,j}) = p_{j,1} x_{i,j}^1 + p_{j,2} x_{i,j}^2 + \cdots + p_{j,D_j} x_{i,j}^{D_j} \tag{2}$$

[4] In MCDA, $\mathbf{x}_i$ is called an alternative with n criteria/attributes.

[5] In GAM, $v(\cdot)$ is called *shape function* and $F(\cdot)$ is called *link function*.

where $p_{j,d} \in \mathbb{R}, d = \{1, 2, \cdots, D_j\}$ is the coefficient of the d-th degree and D_j is the highest order of degree on the j-th attribute.

The motivations using Eq.(2) as a marginal value function are derived from two facets. First, we enhance the expressiveness of the preference model to capture non-monotonic preferences. For example, piecewise linear or monotonic polynomial functions fail to restore all information in a larger learning set (Sobrie et al. 2018). The nonlinearity and non-monotonicity of Eq.(2) can better fit complex relationships between attributes and the outcome, leading to a better model performance. Second, while analyzing human behavior, it is critical to examine the trade-offs or marginal rates of substitution in economics and management studies. A non-derivative value function, for instance the boosted bagged trees model in Lou et al. (2012), cannot capture the inflexion point where the marginal rate of substitution grows or diminishes more quickly (Keeney and Raiffa 1993). A model exploiting human behaviors seems convincing and has more managerial meaning for the DM in management scenarios.

3.2. Neural network-based MCDA.

Full complexity models perform well on many machine learning tasks because they can model both the nonlinearity and the interactions between attributes. An additive model like Eq.(1) does not model any interactions between attributes. Therefore, we propose a neural network-based multiple criteria decision aiding (NN-MCDA) model in the following form

$$U(\mathbf{x}_i) = \alpha \sum\nolimits_{j=1}^{n} w_j v_j(x_{i,j}) + (1-\alpha) f(x_{i,1}, x_{i,2}, \cdots, x_{i,n}) \tag{3}$$

$$= \alpha F(\mathbf{x}_i) + (1-\alpha) f(x_{i,1}, x_{i,2}, \cdots, x_{i,n}) \tag{4}$$

where $U(\mathbf{x}_i)$ is the *global score* of $\mathbf{x}_i$, $f(\cdot)$ is a latent function of all attributes, and $\alpha \in [0, 1]$ is a trade-off coefficient. Eq.(4) describes (a) a regression model if $U(\cdot)$ is the identity, and (b) a classification model if $U(\cdot)$ is the logistic function of the identity. $f(\cdot)$ is used to capture the high-order interrelations between attributes in the model. We can use any complexity models to fit $f(\cdot)$ for better performance, for instance we use a MLP in this paper (Rosenblatt 1958). While using an MLP form of $f(\cdot)$, it is not transparent, meaning that we do not know the exact structure of $f(\cdot)$. Since we have the $F(\cdot)$ to capture the explainable form of the marginal value functions, the non-transparent $f(\cdot)$ describes the

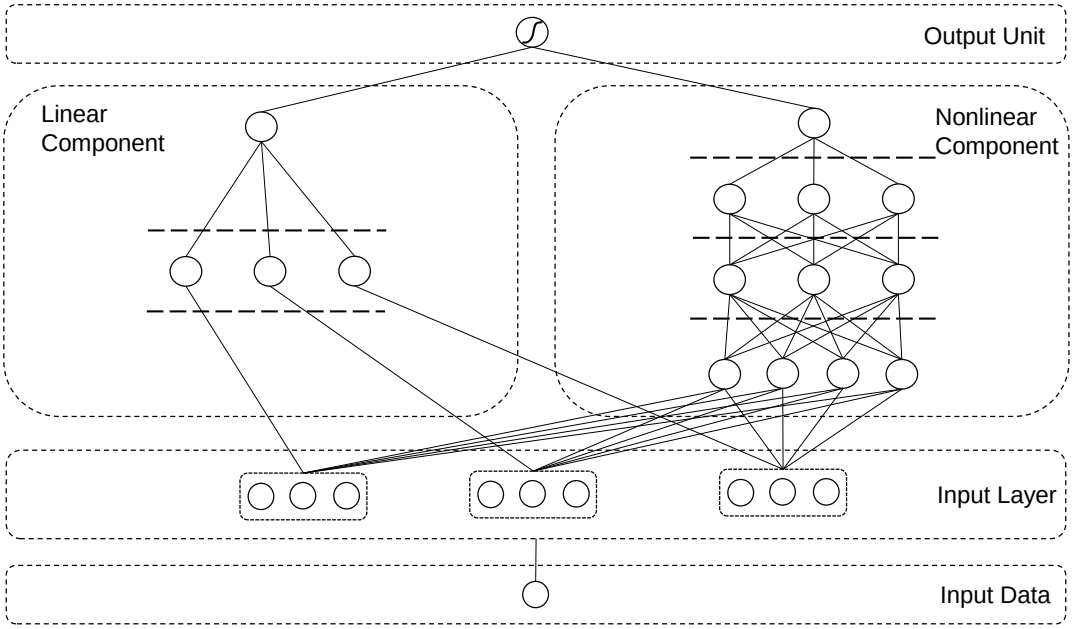

Figure 3 The framework for NN-MCDA.

complex patterns that are not readily useful to the DM. Coefficient α balances the trade-off between $F(\cdot)$ and $f(\cdot)$. If α is close to 1, the model tends to be in a simple additive MCDA form. If α is close to 0, we obtain a full complexity model.

The utilized joint training process is shown in Figure 3. The input attribute vectors should be transformed into polynomial forms, i.e., $\Phi(\mathbf{x}_i) = (\phi(x_{i,1}), \cdots, \phi(x_{i,j}), \cdots, \phi(x_{i,n}))^T = (x_{i,1}^1, \cdots, x_{i,1}^{D_1}, x_{i,2}^1, \cdots, x_{i,2}^{D_2}, \cdots, x_{i,n}^1, \cdots, x_{i,n}^{D_n})^T$. In the input layer, a single-layer network without any activation functions is provided to reconstruct Eq.(1). It has $\sum_{j=1}^{m} D_j$ units and the weight for each unit corresponds to a particular $p_{j,d}$. We denote the output of the linear component with z_i^{linear}, and

$$z_i^{linear} = (w_1, \cdots, w_j, \cdots, w_n) \begin{pmatrix} \mathbf{p}_1^T \phi(x_{i,1}) \\ \vdots \\ \mathbf{p}_j^T \phi(x_{i,j}) \\ \vdots \\ \mathbf{p}_m^T \phi(x_{i,n}) \end{pmatrix} = \mathbf{w}^T \mathbf{P}_i \tag{5}$$

where $\mathbf{p}_j = (p_{j,1}, p_{j,2}, \cdots, p_{j,D_j})^T$ is the vector of coefficients in the j-th polynomial marginal value function, $\mathbf{w}$ is the vector of weights of attributes, and $\mathbf{P}_i$ contains marginal values of i-th attribute vector. Note that, Eq.(5) is actually a specific case of Eq.(1). In Eq.(1), the

marginal value functions $v(\cdot)$ can be in *any* shapes (e.g. piece-wise linear). However, in this study, we allow them to be in a polynomial form in Eq.(2). Thus, $F(\cdot)$ is a generalization of z_i^{linear}.

The nonlinear component is a standard MLP. It is used to learn high-order correlations between attributes. Similarly, by summing every D_j units we can obtain a marginal value on the j-th attribute. For activation functions, we opt for Rectifier (ReLU), which is is the most commonly used activation function in neural networks (Glorot et al. 2011). We can also use other activation functions such as Sigmoid and TanH functions. An L-layer MLP is defined as:

$$\begin{aligned} \mathbf{z}_1(\mathbf{x}_i) &= \Phi(\mathbf{x}_i), \\ \mathbf{z}_2(\mathbf{x}_i) &= a_1(\mathbf{W}_1^T \mathbf{z}_1 + \mathbf{b}_1), \\ &\cdots \\ \mathbf{z}_L(\mathbf{x}_i) &= a_{L-1}(\mathbf{W}_{L-1}^T \mathbf{z}_{L-1} + \mathbf{b}_{L-1}), \\ z_i^{nonlinear} &= \mathbf{h}^T \mathbf{z}_L(\mathbf{x}_i), \end{aligned} \tag{6}$$

where $\mathbf{W}_l$, $\mathbf{b}_l$ and a_l denote the weight matrix, bias vector and activation function for the l-th layer, respectively. The input of the MLP model is the same as the input for the linear part, i.e., $\Phi(\mathbf{x}_i)$.

The output is the probability of $y_i = 1$, we have

$$P(\hat{y}_i = 1|\mathbf{x}_i) = \sigma(\alpha z_i^{linear} + (1-\alpha) z_i^{nonlinear}) \tag{7}$$

where $\sigma(\cdot)$ is a sigmoid function. To estimate the parameters, we minimize the mean square error (MSE):

$$MSE = \frac{1}{N} \sum\nolimits_{i=1}^{N} (P(\hat{y}_i = 1|\mathbf{x}_i) - y_i)^2 \tag{8}$$

We can adopt a variety of optimization methods to minimize Eq.(8), such as Stochastic Gradient Descent (SGD), Adaptive Gradient Algorithm (Adagrad) and Adaptive Moment Estimation (Adam). Please refer to Le et al. (2011) for details of the optimization procedure. The interpretability of the model refers to the capacity in developing marginal value functions, which capture the relationship between individual attributes and prediction. With the proposed model, the DM can know what attributes are more important for the prediction, what values of an attribute are positively or negatively associated to the prediction, and where the convexity and concavity of the function are changed.

3.3. Application to multiple criteria ranking problems.

In this subsection, we will show how to apply NN-MCDA to traditional multiple criteria ranking problems where alternatives are ranked based on the DM's preference. In this paper, alternatives are represented as attribute vectors.

Let $\mathbf{x}_i \succsim \mathbf{x}_k$ denote that an attribute vector $\mathbf{x}_i$ is at least as good as $\mathbf{x}_k$, and $\mathbf{x}_i \succ \mathbf{x}_k$ denote that $\mathbf{x}_i$ is better than $\mathbf{x}_k$. Note that the symbol '$\succsim$'(or '$\succ$') does not necessarily require that each element in $\mathbf{x}_i$ is at least as good as (or better than) that in $\mathbf{x}_k$. It actually indicates that one alternative is at least as good as (or better than) another one based on the DM's judgment. For each pair $(\mathbf{x}_i, \mathbf{x}_k)$, we define $y_{i,k}$ as follows:

$$y_{i,k} = \begin{cases} 1, \text{ if } U(\mathbf{x}_i) \geq U(\mathbf{x}_k), \\ 0, \text{ if } U(\mathbf{x}_k) > U(\mathbf{x}_i), \end{cases} \tag{9}$$

and the difference between global scores of $\mathbf{x}_i$ and $\mathbf{x}_k$ is:

$$\begin{aligned} U(\mathbf{x}_i) - U(\mathbf{x}_k) &= \alpha \sum\nolimits_{j=1}^{n} w_j v_j(x_{i,j}) + (1-\alpha) f(\mathbf{x}_i) - [\alpha \sum\nolimits_{j=1}^{n} w_j v_j(x_{k,j}) + (1-\alpha) f(\mathbf{x}_k)] \\ &= \alpha \sum\nolimits_{j=1}^{n} w_j (v_j(x_{i,j}) - v_j(x_{k,j})) + (1-\alpha)[f(\mathbf{x}_i) - f(\mathbf{x}_k)] \\ &= \alpha \sum\nolimits_{j=1}^{n} w_j \sum\nolimits_{d=1}^{D_j} p_{j,d}(x_{i,j}^d - x_{k,j}^d) + (1-\alpha)[f(\mathbf{x}_i) - f(\mathbf{x}_k)] \\ &= \alpha \sum\nolimits_{j=1}^{n} w_j (\mathbf{p}_j^T \times \phi(x_{i,j}, x_{k,j})) + (1-\alpha)\Theta(\Phi(\mathbf{x}_i, \mathbf{x}_k)) \\ &= \alpha \mathbf{w}^T (\mathbf{P}_i - \mathbf{P}_k) + (1-\alpha)\Theta(\Phi(\mathbf{x}_i, \mathbf{x}_k)) \end{aligned}$$

where $\phi(x_{i,j}, x_{k,j}) = (x_{i,j}^1 - x_{k,j}^1, x_{i,j}^2 - x_{k,j}^2, \cdots, x_{i,j}^{D_j} - x_{k,j}^{D_j})^T$ and $\mathbf{p}_j = (p_{j,1}, p_{j,2}, \cdots, p_{j,D_j})^T$. Let $\Phi(\mathbf{x}_i, \mathbf{x}_k)$ be the aggregated vector for $\phi(x_{i,j}, x_{k,j})$:

$$\Phi(\mathbf{x}_i, \mathbf{x}_k) = \left(\underbrace{x_{i,1}^1 - x_{k,1}^1, \ldots x_{i,1}^{D_1} - x_{k,1}^{D_1}}_{D_1}, \underbrace{x_{i,2}^1 - x_{k,2}^1, \ldots, x_{i,2}^{D_2} - x_{k,2}^{D_2}}_{D_2}, \underbrace{x_{i,3}^1 - x_{k,3}^1, \ldots}_{D_3}, \ldots, \underbrace{\ldots, x_{i,n}^{D_n} - x_{k,n}^{D_n}}_{D_n} \right)^T$$

and $\Theta(\Phi(\mathbf{x}_i, \mathbf{x}_k))$ be a function of $\Phi(\mathbf{x}_i, \mathbf{x}_k)$. We fit $\Theta(\cdot)$ function to approximate the value of $f(\mathbf{x}_i) - f(\mathbf{x}_k)$. Note that in some decision problems, the attribute weights $w_j, j = 1, \cdots, n$ in Eq.(3) are normalized to $[0, 1]$ and $\sum_{j=1}^{n} w_j = 1$, which are useful for interpreting the trade-offs between attributes[6]. To address this issue, we apply the following transformation:

[6] Note that the trade-off between attributes is similar to attribute importance, but the trade-off emphasizes that assigning more weight to an attribute would decrease other attributes. That usually leads to a situation where some attributes have almost no effects on the predictions, which is unexpected because the selected attributes are often summarized based on DM's prior knowledge and their requirements. In this regard, we tend to train our model without normalization but provide normalized weights to evaluate the trade-offs between attributes (Liu et al. 2019). Moreover, there are few minor differences on performances when using normalized weights or not.

- For each attribute $g_j, j = 1, \cdots, n$, the normalized weight is $w'_j = \frac{w_j}{\sum_{j=1}^{n} w_j}$.
- The new global score is $U'(\mathbf{x}_i) = \alpha \sum_{j=1}^{n} w'_j v_j(x_{i,j}) + \frac{(1-\alpha)}{\sum_{j=1}^{n} w_j} f(x_{i,1}, x_{i,2}, \cdots, x_{i,n})$. Moreover, the ordinal relations among all attribute vectors are preserved since $U'(\mathbf{x}_i) - U'(\mathbf{x}_k) = \frac{U(\mathbf{x}_i) - U(\mathbf{x}_k)}{\sum_{j=1}^{n} w_j}$ and $U(\mathbf{x}_i) \geq U(\mathbf{x}_j) \Leftrightarrow U'(\mathbf{x}_i) \geq U'(\mathbf{x}_j)$.

Given the input data $\mathcal{D} = \{(\Phi(\mathbf{x}_i, \mathbf{x}_k), y_{i,k})\}$, instead of mathematical programming, we can now use the machine learning scheme in section 3.2 to infer the preference model and rank other attribute vectors. The output $\hat{y}_{i,k} = \sigma(U(\mathbf{x}_i) - U(\mathbf{x}_k))$ is the probability that $\mathbf{x}_i$ is at least as good as $\mathbf{x}_k$. We can pre-define two thresholds η^1 and η^2, where $0 \leq \eta^1 \leq \eta^2 \leq 1$. If $\eta^2 \leq \hat{y}_{i,k}$, then $\mathbf{x}_i \succ \mathbf{x}_k$, and if $\eta^1 \leq \hat{y}_{i,k} \leq \eta^2$, then $\mathbf{x}_i \sim \mathbf{x}_k$ and otherwise, $\mathbf{x}_k \succ \mathbf{x}_i$. If we use the normalized weights, since the probability $\hat{y}'_{i,k} = \sigma(U'(\mathbf{x}_i) - U'(\mathbf{x}_k))$ is transformed nonlinearly, the pre-defined thresholds should also be transformed as follows, $\eta^1_{i,k} = \frac{\hat{y}'_{i,k}}{y_{i,k}} \eta^1, \eta^2_{i,k} = \frac{\hat{y}'_{i,k}}{y_{i,k}} \eta^2$, to preserve the ordinal relations. In this way, the traditional multiple criteria ranking approaches can handle larger datasets and obtain smoother and more flexible marginal value functions to assist the DM. We present the simulation results in Section 4.1 and the results using real datasets in Section 4.2.

3.4. Usefulness of the proposed framework in decision making.

As we introduce MCDA into machine learning, the main objective is shifted from achieving the best predictive performance to facilitating the DM in gaining insights into the characteristics of the decision making process and the interpretations of the results (Doumpos and Zopounidis 2011). Once the marginal value functions are obtained by the proposed NN-MCDA framework, we can further analyze the DM's preference from the following perspectives.

First, the attribute importance usually has a long-tail distribution, with a few of them being very important and the majority of them being less important (Caruana et al. 2015). The characteristics of the marginal value functions can reveal the importance of the corresponding attribute. If a marginal value function is close to 0 for the whole scale of the attribute values, it indicates that the attribute is either not important to the DM or the characteristic of the marginal value function is wrongly captured, because the change of this attribute has little influence on the predictions. When this is the case, we need to interact with the DM to determine whether we preserve this attribute or calibrate the model. In this regard, the proposed framework can perform model selection and modification (similar to statistical approaches like LASSO). For example, while predicting if a patient has the

flu, the marginal value function of "room humidity" is in a shape like the marginal value function 1 in Figure 1, it is possible that "room humidity" has little contribution to the flu. However, whether we abandon it should be determined by a physician.

Second, the increasing and decreasing tendencies of the marginal value function curves reveal the change of the DM's non-monotonic preference. We focus on the monotonicity inflexion points because they can determine that to what attribute values, the DM is more sensitive. Moreover, if we partition the marginal value function curve by these points, we can discretize the continuous attribute into smaller ranges in which the DM's preference is monotonic. Such smaller intervals are useful for personalization (e.g. customer segmentation) and strategy-making tasks in management. For example, while evaluating the company's performance, if the marginal value function of "cash to total assets ratio" is like the second function in Figure 1, we can learn that a company with a very small or large "cash to total assets ratio" is in a bad condition. Companies with a large ratio are suggested to use the cash to do more investigations, whereas companies with a small ratio are suggested to save general expenses so that more cash can be used in new investigations.

Third, Since the marginal value function returns a "score" that is added to the global value, it is crucial to determine whether the attribute positively or negatively contributes to the outcome. If a marginal value function is above/below *zero*, the corresponding attribute is positively/negatively associated with the prediction. The marginal value function can capture the sign change (if any) of an attribute's contribution and provide the DM an exact attribute value where the sign changes. This is more informative than the statistical models that only provide a fixed coefficient representing either positive or negative effect of the attribute. For example, when predicting the risk of depression among adults, the marginal value function of "age" may has a shape similar to the third function in Figure 1 (please also refer to Figure 18, which is drawn from the real-data). The shape of this curve indicates that the risk of depression does not increase while aging if the adult is younger than a threshold. The risk will increase once the adult is older than that threshold (the threshold is 71.58 in the real data introduced in section 4.3). Statistical models, on the other hand, can only conclude that age has either a negative or positive effect on the depression risk. We need to segment the adults to pre-defined age groups to capture such sign change effect.

Fourth,the concavity (and convexity) of the marginal value function can directly reflect the changing rate of the DM's preference. Such information is important to both economics and marketing problems. For example, if the consumer's preference to "discount rate" is in a same shape of marginal value function 4 in Figure 1, it implies that at the beginning, along with the increase of the discount rate, the consumer's utility (propensity to consume the product) grows more quickly. However, when the discount rate is over a specific value, it gives a signal that the product is possibly of bad quality. Although the consumer's utility still grows, its rate of increase starts to slow down. This provides the DM with a conclusion that keeping the discount rate at a medium level could maximize the profit.

4. Experiments.

To validate the proposed NN-MCDA model, we perform experiments with both synthetic and real datasets. We use *area under the curve* (AUC) of *receiver operating characteristic* (ROC) curve to measure the model performance. In subsection 4.1, three simulation experiments examine (a) the influence of the degree of polynomial on the prediction performance, (b) the influence of the value of α, the trade-off coefficient, on prediction performance, and (c) the goodness of the proposed NN-MCDA approach in fitting the given marginal value functions. In Section 4.2, we first apply the NN-MCDA model to a multiple criteria decision problem where we rank universities based on the employer reputation. Then we predict the risk for geriatric depression with useful interpretations of the risk factors with a higher resolution.

4.1. Simulations.

For brevity, we set equal pre-defined degrees of polynomial for all marginal value functions in the subsequent experiments. We generate three typical synthetic datasets (from the simplest to very complex) as follows:

1. Uniformly draw N attribute vectors with n attributes whose values are within [0,1].

2. We generate three datasets. (a) For the first dataset $\mathcal{D}_l^n$, all n attributes have equal importance and the actual marginal value functions are identity functions. The global score for each attribute vector is a linear summation of n attribute values without any attribute interactions and an additional noise term that is in a standard normal distribution; (b) The second dataset $\mathcal{D}^n_{polynomial-3}$ randomly generates 3-degree polynomials marginal value functions for n attributes, and the global score is the summation of marginal values,

Table 1 **Parameter settings for simulation experiment I.**

	N	# Comparisons	Training size	Pre-defined D_j	# Attributes
Setting	250	$\frac{250*249}{2}$	0.5, 0.6, 0.7, 0.8, 0.9	1, 2, 3, 5, 10	3, 5

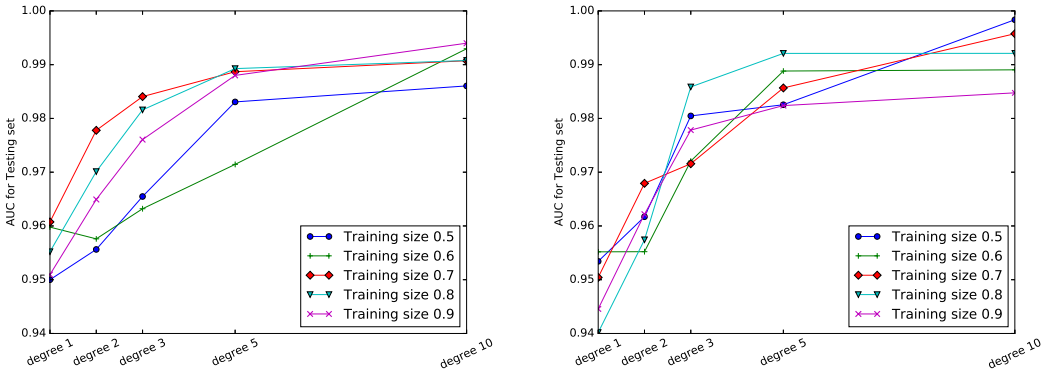

Figure 4 **The average AUC for the testing set using training set with different sizes from datasets $\mathcal{D}_l^3$ and $\mathcal{D}_l^5$.**

all $\binom{n}{2}$ attribute interactions and a standard normal noise term. (c) The third dataset $\mathcal{D}^n_{polynomial-15}$ is extremely complex. The global score is the summation of n 15-degree polynomial marginal values, all possible attribute interactions (pairwise, triple-wise and higher interactions) and a standard normal noise term.

3. We compare global scores between each pair of attribute vectors. If $U(\mathbf{x}_i) - U(\mathbf{x}_k) \geq 0$, then $y_{i,k} = 1$, otherwise, $y_{i,k} = 0$. Note that the actual input is the transformed attribute vector.

4.1.1. Experiment I: Relationship between degree of polynomial and model performance The first simulated experiment aims at exploring the relationship between the pre-defined degree of polynomial and AUC. The parameters used in the experiment is shown in Table 1. For each setting, we iteratively repeat the experiment for 10 times and record the averaged AUC. In this experiment, the numbers of iterations are determined using fivefold cross-validation: We partition the training set into five sets and set aside one of them as a validation set. We then train the model using the other four partitions and use the validation set to check the convergence. This procedure is repeated five times and the averaged number of iterations is used to train the final model with the whole dataset (Lou et al. 2012).

Figures 4, 5 and 6 report the averaged AUC for the testing set with different training sizes using the three synthetic datasets. Though there is no obvious relationships between the training sizes and the model performance, we find two interesting patterns. First, higher

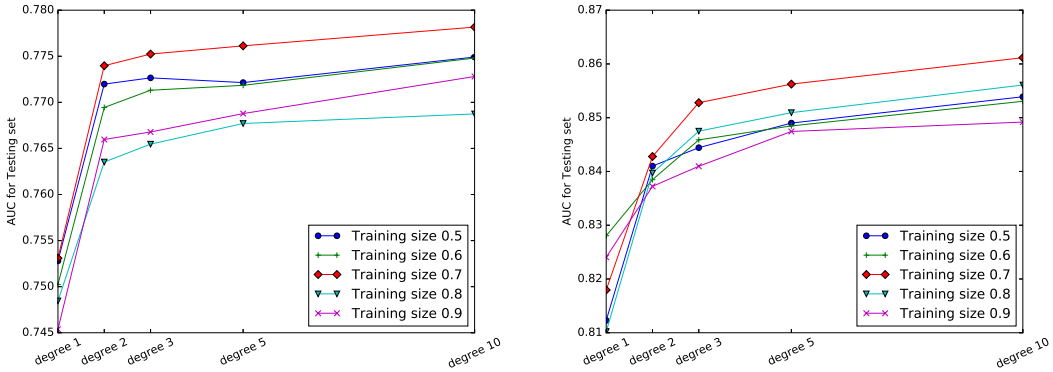

Figure 5 **The average AUC for the testing set using training set with different sizes from datasets $\mathcal{D}^3_{polynomial-3}$ and $\mathcal{D}^5_{polynomial-3}$.**

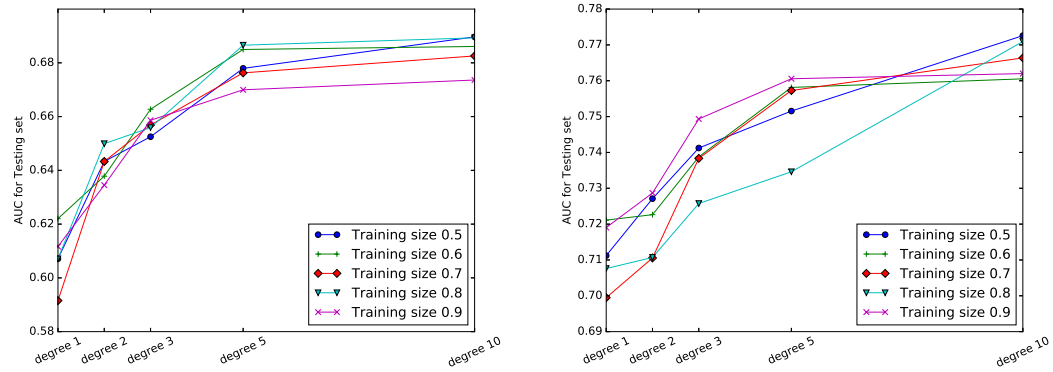

Figure 6 **The average AUC for the testing set using training set with different sizes from datasets $\mathcal{D}^3_{polynomial-15}$ and $\mathcal{D}^5_{polynomial-15}$.**

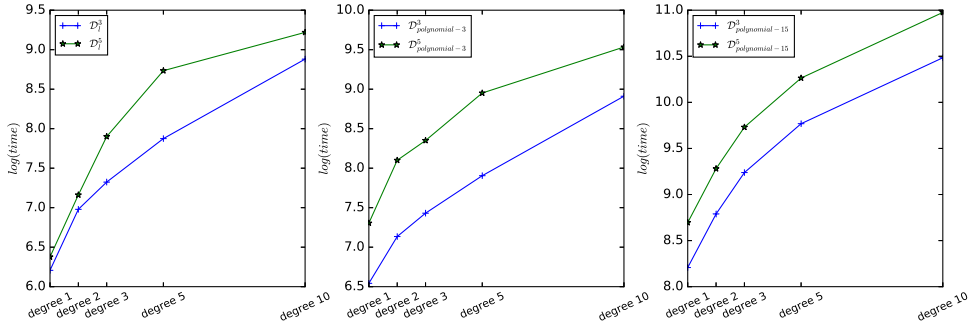

Figure 7 **The average computational time for different datasets.**

Note. The recoded computational times in the figure are the averaged time for training all five training sizes given a pre-defined degree of polynomials.

pre-defined degrees of polynomials can lead to higher accuracy when convergence. That results from the ability of the underlying model to capture more complicated nonlinearity.

Table 2 The average AUCs ± one deviation on dataset $\mathcal{D}_l^3$ and $\mathcal{D}_l^5$ using different machine learning algorithms.

	n=3					n=5				
Training size	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
NN-MCDA-D1	0.949±0.023	0.959±0.021	0.960±0.011	0.955±0.009	0.951±0.005	0.953±0.021	0.955±0.014	0.951±0.019	0.941±0.013	0.945±0.012
NN-MCDA-D2	0.956±0.022	0.958±0.027	0.978±0.019	0.971±0.007	0.965±0.003	0.962±0.018	0.955±0.021	0.968±0.009	0.957±0.022	0.962±0.018
NN-MCDA-D3	0.965±0.031	0.963±0.024	0.984±0.019	0.981±0.004	0.976±0.003	0.980±0.019	0.972±0.019	0.972±0.021	0.986±0.013	0.978±0.010
NN-MCDA-D5	0.983±0.026	0.971±0.020	0.988±0.019	0.989±0.005	0.988±0.001	0.983±0.028	0.989±0.018	0.876±0.037	0.992±0.009	0.982±0.010
NN-MCDA-D10	0.986±0.011	0.993±0.010	0.991±0.009	0.991±0.002	0.994±0.001	0.998±0.013	0.989±0.029	0.996±0.009	0.992±0.007	0.985±0.008
MLP-D1	0.969±0.001	0.951±0.005	0.960±0.000	0.957±0.000	0.998±0.001	0.970±0.001	0.969±0.002	0.977±0.001	0.981±0.001	0.966±0.000
MLP-D2	0.970±0.002	0.969±0.003	0.973±0.003	0.970±0.001	0.999±0.002	0.979±0.001	0.980±0.000	0.989±0.001	0.988±0.001	0.988±0.001
MLP-D3	0.992±0.001	0.990±0.001	0.987±0.002	0.989±0.001	0.973±0.001	0.987±0.002	0.991±0.001	0.991±0.001	0.995±0.000	0.996±0.000
MLP-D5	0.999±0.001	0.993±0.000	0.990±0.001	0.998±0.000	0.999±0.001	0.995±0.001	0.996±0.001	0.995±0.002	0.996±0.000	0.996±0.000
MLP-D10	0.999±0.002	0.997±0.001	0.995±0.000	0.999±0.001	0.999±0.001	0.999±0.001	0.999±0.001	0.998±0.001	0.998±0.001	0.999±0.000
SVM-Linear	0.998±0.001	0.998±0.001	0.998±0.001	0.998±0.000	0.998±0.000	0.997±0.002	0.998±0.003	0.997±0.001	0.997±0.000	0.997±0.000
SVM-RBF	0.998±0.001	0.997±0.000	0.998±0.001	0.998±0.000	0.999±0.001	0.995±0.001	0.996±0.000	0.996±0.003	0.996±0.001	0.996±0.001
SVM-D3poly	0.967±0.002	0.969±0.001	0.962±0.000	0.971±0.001	0.973±0.000	0.987±0.003	0.991±0.000	0.986±0.001	0.988±0.002	0.989±0.001
GAM-D3	0.983±0.011	0.983±0.017	0.983±0.009	0.982±0.003	0.985±0.002	0.986±0.010	0.985±0.011	0.984±0.010	0.986±0.009	0.984±0.010
GAM-D10	0.985±0.010	0.984±0.018	0.984±0.010	0.983±0.002	0.985±0.001	0.987±0.008	0.985±0.009	0.985±0.011	0.986±0.010	0.986±0.011
DeciTr-MaxDep6	0.842±0.001	0.842±0.003	0.843±0.001	0.835±0.001	0.831±0.000	0.931±0.001	0.927±0.000	0.937±0.000	0.933±0.000	0.938±0.000
DeciTr-MaxDep10	0.896±0.002	0.893±0.000	0.890±0.002	0.897±0.001	0.898±0.000	0.965±0.001	0.965±0.000	0.971±0.001	0.971±0.000	0.967±0.000
DeciTr-MaxDep20	0.900±0.003	0.898±0.001	0.895±0.000	0.908±0.000	0.903±0.000	0.966±0.001	0.966±0.001	0.972±0.001	0.974±0.001	0.970±0.000
PLR-D1	0.899±0.014	0.903±0.022	0.901±0.021	0.900±0.011	0.901±0.000	0.903±0.011	0.902±0.019	0.900±0.016	0.910±0.018	0.911±0.009
PLR-D2	0.910±0.017	0.913±0.023	0.928±0.012	0.923±0.016	0.922±0.001	0.923±0.017	0.931±0.011	0.929±0.019	0.931±0.011	0.936±0.016
PLR-D3	0.945±0.020	0.933±0.017	0.945±0.014	0.941±0.004	0.914±0.001	0.941±0.010	0.952±0.009	0.946±0.014	0.956±0.012	0.944±0.010
PLR-D5	0.950±0.018	0.941±0.023	0.955±0.020	0.949±0.003	0.925±0.000	0.958±0.011	0.959±0.012	0.961±0.012	0.964±0.010	0.961±0.011
PLR-D10	0.959±0.011	0.957±0.009	0.960±0.008	0.955±0.010	0.933±0.000	0.963±0.013	0.970±0.023	0.972±0.014	0.977±0.009	0.973±0.009
Mean	0.957±0.010	0.955±0.011	0.955±0.008	0.958±0.004	0.957±0.001	0.970±0.011	0.958±0.009	0.957 ±0.009	0.974±0.007	0.972±0.006

However, higher degrees of polynomials usually require more iterations to converge. More specifically, we depict the averaged computational time for each training process in Figure 7. Apparently, while increasing the model complexity, for example, using higher degrees of polynomial marginal value function and considering more attributes, the average computational time to converge also increases almost linearly. Another interesting pattern is that the shape of the AUC curves (Figures 4, 5 and 6) can fit a concave function in general. When the degree increases, the AUC improvement (over the model with the immediate smaller degree) is becoming smaller. For example, the improvement is more obvious if we change the pre-defined degree from 1 to 3 than that if we change from 3 to 5 and from 5 to 10. The improvement diminishes quickly along with the increase of pre-defined degree of polynomials. The greatest AUC improvement happens if we increase the degree to 3, while the improvement resulted from further increasing the degree to 5 and 10 is slim. The results suggest that it is not necessary to set a very large D_j for the seek of minor improvement because the computational cost increases much faster when D_j increases. Generally, we believe that a polynomial of 3 degree is sufficient to capture the characteristics for all the three datasets. Higher degrees of polynomials have a risk for over-fitting and obviously cost more computational time, but contribute little to accuracy.

We also compare the proposed NN-MCDA with baseline machine learning models, including the standard MLP, polynomial linear regression (PLR) with 1, 2, 3, 5 and 10 degrees, GAM with 10 splines that are in 3 and 10 degree of polynomials, SVMs with linear, radial basis function (RBF) and polynomial kernels, and single decision tree (DeciTr)

Table 3 The average AUCs ± one deviation on dataset $\mathcal{D}^3_{polynomial-3}$ and $\mathcal{D}^5_{polynomial-3}$ using different machine learning algorithms.

	n=3					n=5				
Training size	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
NN-MCDA-D1	0.753±0.018	0.750±0.023	0.753±0.021	0.748±0.010	0.745±0.008	0.812±0.021	0.828±0.038	0.818±0.029	0.810±0.021	0.824±0.015
NN-MCDA-D2	0.772±0.024	0.769±0.021	0.774±0.012	0.764±0.008	0.766±0.009	0.841±0.023	0.838±0.033	0.843±0.021	0.840±0.019	0.837±0.021
NN-MCDA-D3	0.773±0.026	0.771±0.020	0.775±0.021	0.765±0.011	0.767±0.011	0.844±0.031	0.846±0.052	0.853±0.041	0.847±0.037	0.841±0.031
NN-MCDA-D5	0.772±0.018	0.772±0.019	0.776±0.015	0.768±0.011	0.769±0.008	0.849±0.029	0.853±0.027	0.856±0.039	0.851±0.015	0.847±0.011
NN-MCDA-D10	0.775±0.020	0.775±0.021	0.778±0.012	0.769±0.010	0.773±0.006	0.854±0.032	0.853±0.052	0.861±0.025	0.856±0.030	0.849±0.028
MLP-D1	0.787±0.010	0.793±0.002	0.790±0.001	0.791±0.003	0.800±0.004	0.837±0.002	0.835±0.001	0.831±0.002	0.836±0.001	0.833±0.011
MLP-D2	0.790±0.005	0.793±0.001	0.792±0.003	0.797±0.002	0.802±0.008	0.848±0.001	0.851±0.003	0.845±0.004	0.849±0.010	0.847±0.008
MLP-D3	0.791±0.002	0.799±0.002	0.799±0.001	0.801±0.002	0.808±0.002	0.855±0.004	0.855±0.001	0.851±0.000	0.855±0.000	0.857±0.000
MLP-D5	0.793±0.001	0.806±0.003	0.801±0.002	0.805±0.001	0.810±0.006	0.861±0.002	0.856±0.001	0.854±0.003	0.861±0.001	0.860±0.001
MLP-D10	0.795±0.001	0.810±0.002	0.804±0.004	0.810±0.003	0.811±0.009	0.866±0.004	0.864±0.001	0.865±0.000	0.862±0.001	0.862±0.000
SVM-Linear	0.687±0.004	0.683±0.004	0.686±0.004	0.681±0.006	0.678±0.001	0.748±0.003	0.748±0.003	0.755±0.003	0.748±0.007	0.743±0.007
SVM-RBF	0.688±0.003	0.685±0.003	0.686±0.007	0.682±0.005	0.680±0.000	0.748±0.003	0.748±0.002	0.755±0.003	0.748±0.004	0.743±0.006
SVM-D3poly	0.682±0.004	0.680±0.003	0.687±0.006	0.682±0.005	0.676±0.000	0.750±0.002	0.749±0.003	0.756±0.005	0.752±0.004	0.749±0.008
GAM-D3	0.687±0.002	0.684±0.001	0.688±0.002	0.681±0.001	0.681±0.001	0.749±0.011	0.746±0.006	0.752±0.008	0.750±0.007	0.742±0.003
GAM-D10	0.687±0.004	0.684±0.006	0.687±0.009	0.681±0.001	0.680±0.002	0.749±0.010	0.746±0.010	0.752±0.012	0.749±0.011	0.741±0.009
DeciTr-MaxDep6	0.685±0.003	0.678±0.003	0.682±0.004	0.679±0.007	0.676±0.006	0.724±0.003	0.725±0.002	0.734±0.005	0.736±0.006	0.725±0.009
DeciTr-MaxDep10	0.672±0.002	0.669±0.004	0.672±0.003	0.663±0.007	0.665±0.006	0.717±0.002	0.719±0.002	0.721±0.004	0.722±0.002	0.722±0.008
DeciTr-MaxDep20	0.617±0.003	0.623±0.002	0.620±0.005	0.621±0.009	0.630±0.008	0.666±0.003	0.663±0.004	0.672±0.003	0.671±0.005	0.665±0.009
PLR-D1	0.604±0.011	0.609±0.021	0.611±0.014	0.603±0.011	0.610±0.003	0.703±0.039	0.698±0.010	0.700±0.019	0.702±0.010	0.708±0.017
PLR-D2	0.621±0.013	0.632±0.018	0.639±0.013	0.625±0.010	0.621±0.010	0.712±0.010	0.702±0.018	0.710±0.010	0.711±0.015	0.713±0.029
PLR-D3	0.638±0.021	0.644±0.020	0.640±0.009	0.637±0.009	0.635±0.010	0.730±0.011	0.729±0.021	0.721±0.014	0.729±0.029	0.720±0.018
PLR-D5	0.651±0.017	0.657±0.017	0.649±0.019	0.650±0.011	0.650±0.011	0.733±0.011	0.735±0.013	0.729±0.020	0.734±0.016	0.736±0.019
PLR-D10	0.669±0.009	0.667±0.007	0.670±0.011	0.669±0.011	0.668±0.008	0.739±0.021	0.739±0.023	0.732±0.009	0.741±0.009	0.739±0.014
Mean	0.713±0.010	0.708±0.010	0.713±0.009	0.712± 0.006	0.713±0.006	0.780±0.009	0.779±0.014	0.781±0.012	0.775±0.009	0.778±0.012

Table 4 The average AUCs ± one deviation on dataset $\mathcal{D}^3_{polynomial-15}$ and $\mathcal{D}^5_{polynomial-15}$ using different machine learning algorithms.

	n=3					n=5				
Training size	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
NN-MCDA-D1	0.607±0.010	0.622±0.021	0.591±0.009	0.607±0.010	0.612±0.010	0.711±0.024	0.721±0.032	0.699±0.012	0.708±0.017	0.719±0.022
NN-MCDA-D2	0.643±0.008	0.638±0.016	0.643±0.012	0.650±0.008	0.635±0.010	0.727±0.027	0.723±0.029	0.711±0.026	0.711±0.016	0.729±0.021
NN-MCDA-D3	0.653±0.023	0.663±0.018	0.657±0.019	0.656±0.021	0.659±0.013	0.741±0.028	0.739±0.043	0.738±0.037	0.726±0.029	0.749±0.013
NN-MCDA-D5	0.678±0.019	0.685±0.020	0.676±0.005	0.687±0.023	0.670±0.009	0.752±0.035	0.758±0.039	0.757±0.032	0.735±0.035	0.761±0.010
NN-MCDA-D10	0.690±0.023	0.686±0.021	0.683±0.004	0.689±0.009	0.674±0.006	0.773±0.022	0.761±0.048	0.766±0.019	0.771±0.029	0.762±0.031
MLP-D1	0.710±0.001	0.711±0.005	0.710±0.000	0.713±0.000	0.712±0.001	0.790±0.002	0.787±0.004	0.780±0.001	0.779±0.002	0.789±0.003
MLP-D2	0.713±0.003	0.712±0.002	0.714±0.002	0.718±0.003	0.715±0.002	0.794±0.003	0.790±0.003	0.782±0.001	0.783±0.002	0.791±0.001
MLP-D3	0.719±0.004	0.719±0.003	0.720±0.001	0.721±0.002	0.719±0.003	0.799±0.003	0.794±0.002	0.791±0.002	0.788±0.001	0.800±0.003
MLP-D5	0.721±0.002	0.720±0.000	0.724±0.003	0.727±0.003	0.725±0.004	0.803±0.001	0.799±0.005	0.801±0.002	0.794±0.003	0.803±0.001
MLP-D10	0.728±0.003	0.724±0.001	0.729±0.002	0.730±0.002	0.729±0.003	0.809±0.002	0.803±0.001	0.805±0.002	0.806±0.003	0.808±0.002
SVM-Linear	0.579±0.005	0.579±0.002	0.584±0.003	0.572±0.005	0.579±0.006	0.658±0.003	0.659±0.005	0.649±0.004	0.645±0.002	0.658±0.005
SVM-RBF	0.586±0.003	0.584±0.004	0.586±0.002	0.578±0.002	0.589±0.005	0.660±0.001	0.661±0.002	0.654±0.003	0.648±0.003	0.658±0.003
SVM-D3poly	0.574±0.002	0.579±0.006	0.580±0.006	0.572±0.004	0.579±0.005	0.659±0.004	0.662±0.002	0.652±0.002	0.647±0.002	0.655±0.006
GAM-D3	0.579±0.001	0.579±0.002	0.584±0.003	0.572±0.001	0.576±0.003	0.659±0.002	0.659±0.004	0.650±0.003	0.647±0.005	0.656±0.004
GAM-D10	0.583±0.001	0.582±0.003	0.585±0.003	0.579±0.002	0.583±0.005	0.660±0.004	0.659±0.003	0.651±0.003	0.645±0.002	0.657±0.006
DeciTr-MaxDep6	0.583±0.003	0.583±0.003	0.583±0.003	0.581±0.005	0.589±0.009	0.649±0.002	0.650±0.001	0.644±0.004	0.638±0.005	0.651±0.008
DeciTr-MaxDep10	0.565±0.004	0.579±0.002	0.574±0.004	0.570±0.007	0.562±0.008	0.631±0.004	0.635±0.003	0.632±0.005	0.628±0.006	0.638±0.009
DeciTr-MaxDep20	0.536±0.005	0.548±0.006	0.539±0.004	0.541±0.007	0.547±0.008	0.592±0.003	0.583±0.004	0.591±0.002	0.588±0.008	0.582±0.007
PLR-D1	0.560±0.015	0.562±0.026	0.558±0.021	0.559±0.017	0.557±0.011	0.632±0.024	0.630±0.017	0.629±0.019	0.631±0.022	0.633±0.020
PLR-D2	0.562±0.014	0.568±0.021	0.560±0.020	0.560±0.010	0.562±0.017	0.637±0.019	0.631±0.021	0.633±0.011	0.633±0.017	0.635±0.019
PLR-D3	0.569±0.010	0.569±0.019	0.564±0.019	0.565±0.028	0.563±0.028	0.642±0.023	0.637±0.019	0.637±0.023	0.639±0.028	0.639±0.020
PLR-D5	0.571±0.014	0.572±0.023	0.569±0.020	0.569±0.027	0.569±0.027	0.647±0.022	0.642±0.020	0.640±0.029	0.641±0.018	0.641±0.018
PLR-D10	0.578±0.015	0.575±0.012	0.572±0.023	0.575±0.019	0.573±0.019	0.652±0.021	0.649±0.011	0.644±0.016	0.645±0.019	0.647±0.022
Mean	0.621±0.005	0.623±0.010	0.621±0.008	0.621±0.005	0.621±0.005	0.699±0.012	0.697±0.014	0.693±0.011	0.690±0.013	0.698±0.011

models with 6, 10 and 20 maximum depths. Table 2 presents the results for the simplest dataset. All machine learning models, both the interpretable (including NN-MCDA) and full complexity ones, perform well. The performance drops rapidly when we use them to fit the nonlinear and high-order datasets (shown in Tables 3 and 4). Since the proposed NN-MCDA model and MLP can model the nonlinearity and attribute interactions, both of them achieve much higher AUCs as compared to the rest. As expected, although the performance of NN-MCDA is lower than MLP, the difference is relatively small. Both NN-MCDA and MLP outperform other baseline machine learning models significantly. More

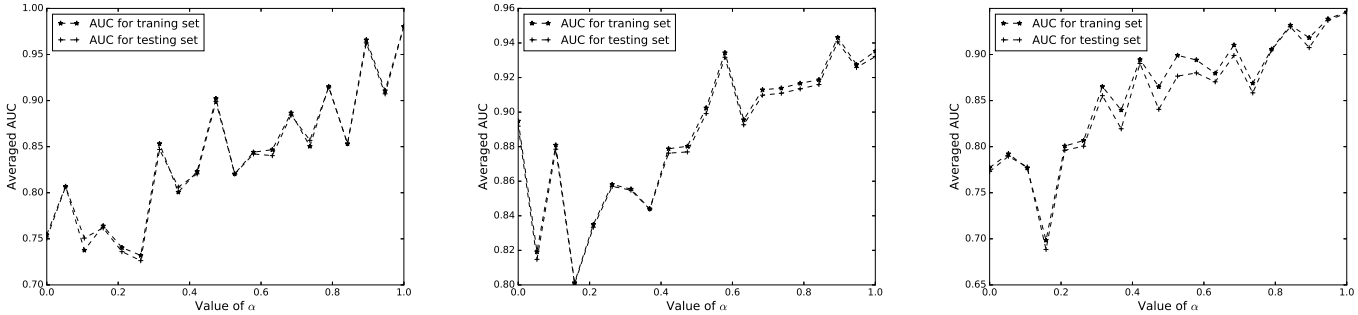

Figure 8 **From the left to right, the impact of α on AUC using $\mathcal{D}_l^3, \mathcal{D}_l^5$, $\mathcal{D}_l^{10}$, respectively.**

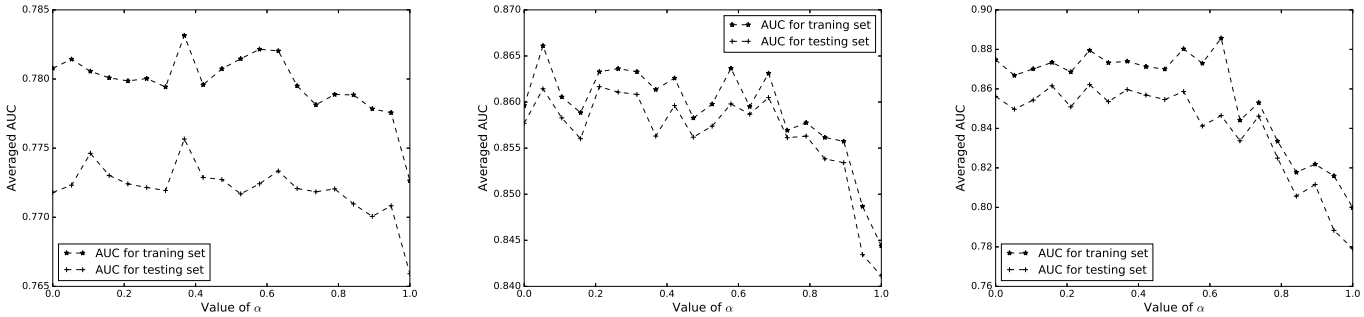

Figure 9 **From the left to right, the impact of α on AUC using $\mathcal{D}^3_{polynomial-3}, \mathcal{D}^5_{polynomial-3}$, $\mathcal{D}^{10}_{polynomial-3}$, respectively.**

specifically, we can observe that GAM, SVM and DeciTr have similar accuracy, which is higher than that of PLR, but lower than NN-MCDA. NN-MCDA's performance is close to the full complexity model, and at the same time, it has strong interpretability (will be shown in the next experiment).

4.1.2. Experiment II: Impact of α on AUC In this section, we focus on assessing how the performance of the NN-MCDA model is affected by the value of α, the weight for the linear component. We evenly sample 20 values within [0, 1] as the pre-defined α. For each fixed α, we train the NN-MCDA model using synthetic datasets introduced in the previous subsection. In this experiment, we use the SGD algorithm to optimize the parameters and the number of iterations are set as 250.

In Figure 8 (results with the linearly generated datasets), though three curves have many monotonicity inflexions, there is a general trend that the greater the value of α, the better the performance. NN-MCDA obtains the best results when α is between 0.8 and 1.0 on these linearly generated datasets. In contrast, the AUC curves have a general decreasing trend when α increases for the nonlinearly generated datasets (Figures 9 and

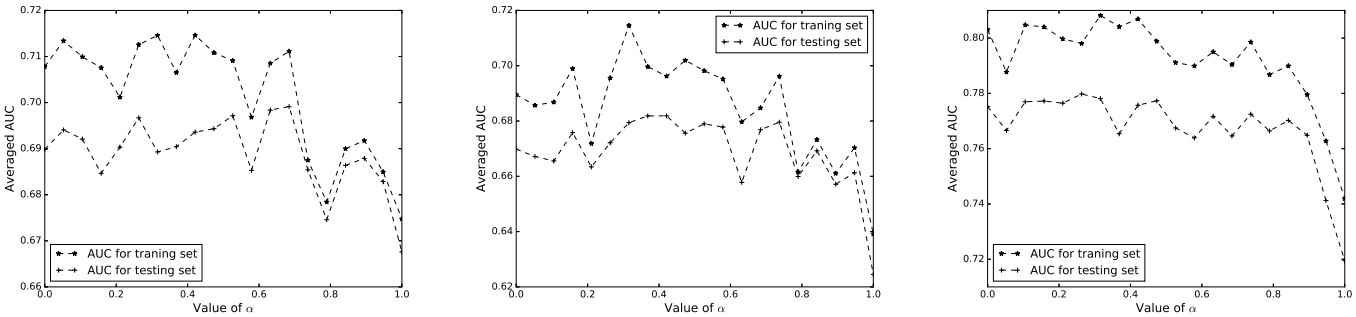

Figure 10 **From the left to right, the impact of α on AUC using $\mathcal{D}^3_{polynomial-15}$, $\mathcal{D}^5_{polynomial-15}$, $\mathcal{D}^{10}_{polynomial-15}$, respectively.**

10). In the extreme cases, where only the nonlinear MLP component (i.e., $\alpha = 0$) or the linear component (i.e.,$\alpha = 1$) works, the model obtains the greatest or smallest average AUC.

Note that the three datasets have very different patterns. Datasets $\mathcal{D}^n_l$ use a set of simple linear marginal value functions whereas $\mathcal{D}^n_{polynomial-3}$ and $\mathcal{D}^n_{polynomial-15}$ simulate more complicated patterns. Theoretically, a full complexity model (MLP) can perfectly capture any patterns in the data at the cost of very large numbers of iterations and data samples for convergence. In practice, we often do not have sufficient data or computational time to achieve the optimal MLP solution. In this simulation experiment (31,125 data points and 250 iterations to fit the model), a pure MLP model (NN-MCDA with $\alpha = 0$) does not always lead to the best outcome. This result indicates that, in real-world managerial decision making, a full complexity model is usually not the best one not only because of the lack of interpretability, but also the limited data and computational resources to optimize the model. It is sensible to allow the model to automatically adjust the trade-off coefficient α to avoid the scenarios where a very complex model is used to fit simple data, or a simple model is used to fit complex data.

4.1.3. Experiment III: performance in fitting actual marginal value functions This experiment studies the ability of the NN-MCDA model to reconstruct the actual marginal value functions. From Experiments I and II, we find that the NN-MCDA model with degree equal to 3 has a good balance between prediction performance and computational cost. Therefore, in this experiment, we generate four typical synthetic models with different complexities. Each hypothetical model has three marginal value functions to estimate. Then, we use the synthetic models to generate datasets with the same attribute vectors (model

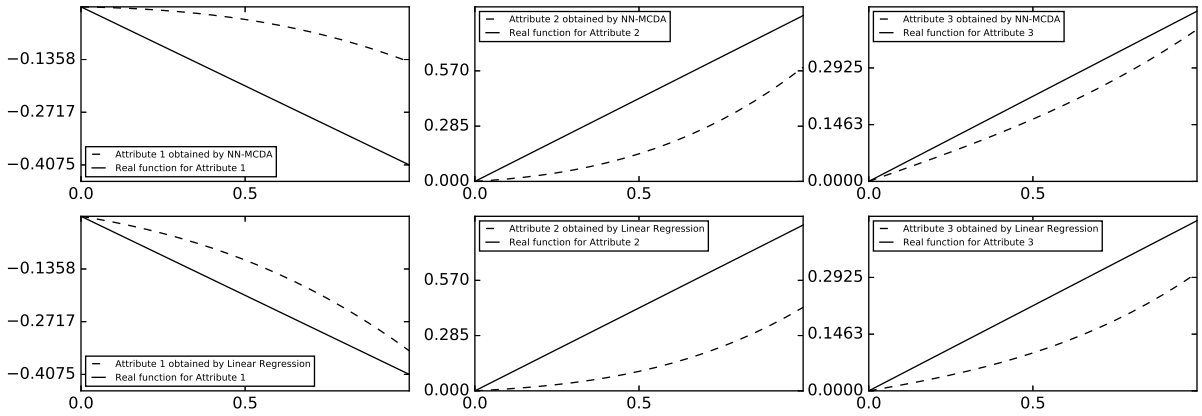

Figure 11 The simulated and actual marginal value functions in model 1.

input). We approximate the marginal value functions using an NN-MCDA model with a degree of 3. We also compare the obtained function with a baseline linear regression model. The four synthetic models (from the simplest to extremely complicated) are described as follows:

- Synthetic model 1: Three linear marginal value functions. The global scores are the linear summation of three marginal values without interactions.
- Synthetic model 2: Three polynomial functions of degree 3. The global scores are the linear summation of three marginal values with all possible pairwise interactions.
- Synthetic model 3: A polynomial function of degree 15, a sigmoid function and an exponential function. The global scores are the linear summation of three marginal values without interactions.
- Synthetic model 4: The Model 3 with pairwise and triple-wise attribute interactions.

Figures 11 to 14 reveal the actual and fitted marginal value functions obtained by the proposed NN-MCDA model and the baseline linear regression model. For a simple model with linear marginal value functions (Synthetic model 1), both the baseline linear regression model and the NN-MCDA can fit the actual functions well. Both models successfully capture the monotonicity of the original marginal value functions. This indicates that the NN-MCDA model is also applicable to simple prediction tasks that do not have attribute interactions or nonlinear associations between attributes and predictions.

When the attribute interactions are considered (Synthetic model 2), the proposed model outperforms the baseline linear regression model. In the first row of Figure 12, the NN-MCDA model captures correct monotonicity changes of all three actual marginal value

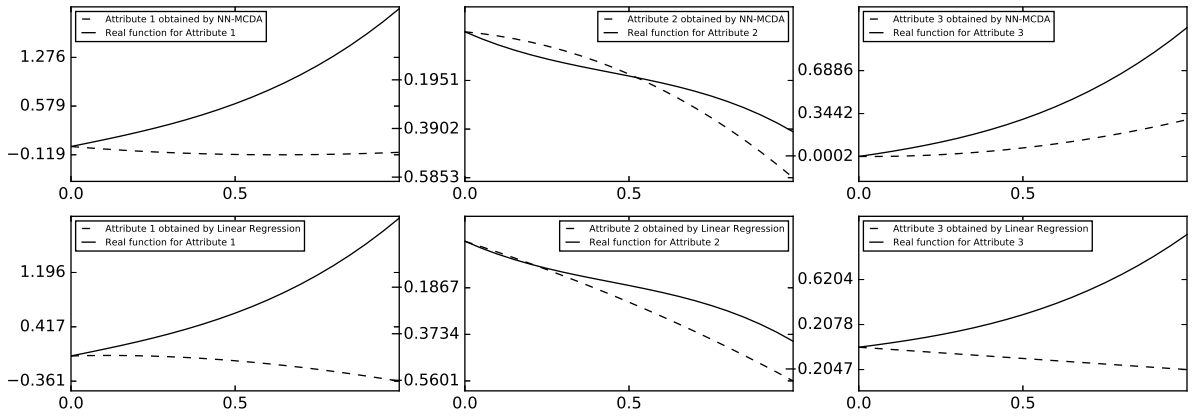

Figure 12 The simulated and actual marginal value functions in model 2.

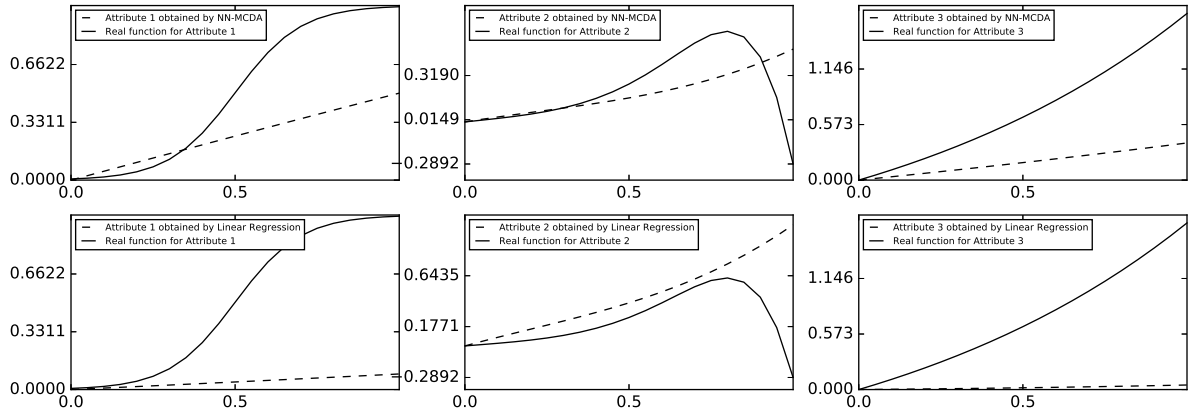

Figure 13 The simulated and actual marginal value functions in model 3.

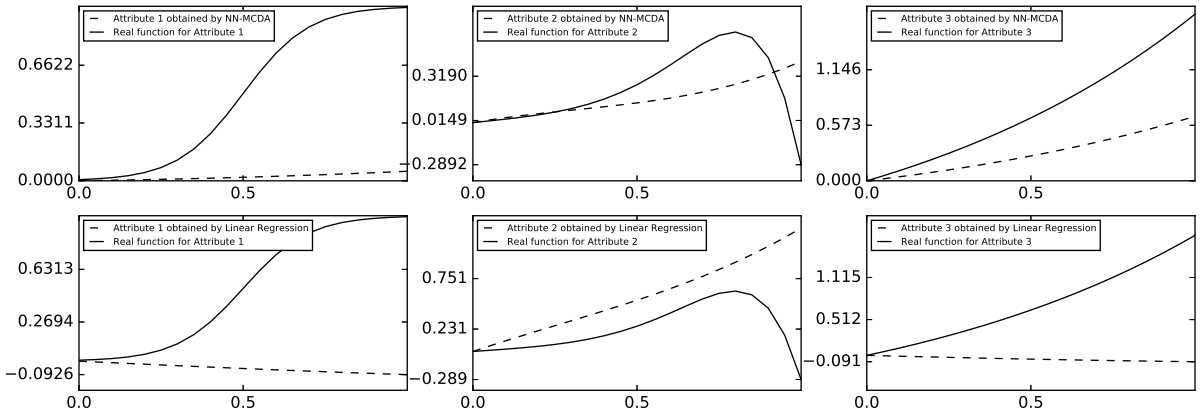

Figure 14 The simulated and actual marginal value functions in model 4.

functions. Moreover, for the first and third attributes, it captures the concavity of the original functions. While linear regression model gives opposite monotonicity of the actual functions except for the second attribute. In addition, the linear regression model does not capture the concavity of the actual functions. The bad performance of the linear regression model is due to the fact that it can not capture the interactions between attributes, which often brings distorted interpretability.

For the model with more complex marginal value functions (Synthetic model 3), the linear regression model performs rather bad. In Figure 13, the first and third attributes have very negligible impact on the prediction in the fitted linear regression model. The NN-MCDA model, on the other hand, correctly captures the main characteristics of the three attributes. Both models failed to capture the inflexion point for the second attribute. These results demonstrate that a low-degree NN-MCDA model (3-degree in this study) can fit both lower and higher degree marginal value functions, because of the nonlinear component helps deal with the complexities that cannot be captured by the predefined linear component. In Figure 14, if the marginal value function is extremely complex (Synthetic model 4), though low-degree NN-MCDA cannot fully capture the characteristics of the marginal value functions, it still outperforms the baseline linear regression model. However, the real-world DM behaviors are usually not that complex (15-degree in Synthetic model 3 and 4). We will further validate the applicability and generalizability of the proposed NN-MCDA model with real datasets in the following section.

4.2. A multiple criteria ranking problem.

QS world university ranking organization[7] provides five carefully-chosen indicators to measure the universities' capacity in producing the most employable graduates, including employer reputation, employer-student connection, alumni outcomes, partnerships with employers and graduate employment rate. The metric of employer reputation, regarded as a key performance indicator, is based on over 40,000 responses to the QS Employer Survey. In this experiment, we apply the proposed NN-MCDA model to a multiple criteria ranking problem that predicts the employer reputation (human decision) using the other four quantitative indicators[8]. The descriptive statistics are shown in Table 5.

[7] It is an annual publication of university rankings by Quacquarelli Symonds (QS). `https://www.topuniversities.com`

[8] **Employer-student connection** (EC). This indicator involves the number of active presences of employers on a university's campus over the past 12 months. Such presences are in form of providing students with opportunities to

Table 5 The descriptive statistics of data.

	Mean	Std Dev	Minimum	Maximum
Employer-student connection	74.63	11.17	44.80	100.00
Alumni outcomes	55.25	16.74	26.40	100.00
Partnerships with employers	73.27	10.76	46.10	100.00
Graduate employment rate	77.88	7.93	54.00	100.00

Original data is normalized into $[0, 1]$ in the following experiment.

Table 6 Given polynomials of pre-defined degree, the averaged AUC for training and testing sets.

	NN-MCDA	Logistic regression	MLP	GAM (5 splines)
1-degree Training	0.66 (0.8988)	0.61 (1.00)	0.71(0.00)	0.64
1-degree Testing	0.64	0.57	0.70	0.63
2-degree Training	0.69 (0.4969)	0.64 (1.00)	0.73 (0.00)	0.67
2-degree Testing	0.68	0.63	0.71	0.66
3-degree Training	0.69 (0.4351)	0.65 (1.00)	0.74 (0.00)	0.68
3-degree Testing	0.68	0.65	0.73	0.66

The value in parentheses is α when convergence. There are minor differences between averaged AUC when pre-defined degrees are 2 and 3. Using polynomial of degree 2 is sufficient because more complex model does not significantly increase AUC.

There are $\binom{250}{2}$ pairwise comparisons among 250 universities. To determine the pre-defined degree of polynomials, we respectively set D_j as 1, 2, and 3. We use the same fivefold cross-validation process to fit the model. We record the averaged AUC for both training and testing sets. We also select three baseline models, including a 3-layer MLP, a logistic regression model and a GAM. The results of the average AUC are presented in Table 6. For each pre-defined degree, the full complexity model always obtain the best results whereas the logistic regression model performs the worst. Since NN-MCDA can model attribute interactions, it slightly outperforms the GAM. As interpretable models, we depict the marginal value functions obtained by NN-MCDA, linear model and GAM in Figures 15, 16 and 17, respectively.

In Figures 15 and 16, the vertical axis is the individual attributes contributions to employer reputation. For the attributes *alumni outcomes*, *partnerships with employers* and *graduate employment rate*, the value functions obtained by NN-MCDA and logistic regression model exhibit a monotonically increasing trend, which makes sense based on our

network and acquire information, organizing company presentations or other self-promoting activities, which increase the probability that students have to participate in career-launching internships and research opportunities.
Alumni outcomes (AO). The scores based on the outcomes of a university's graduates produced. A university is successful if its graduates tend to produce more wealth and scientific researches.
Partnerships with employers (PE). The number of citable and transformative researches which are produced by a university collaborating successfully with global companies.
Graduate employment rate (GER). This indicator is essential for understanding how successful universities are at nurturing employability. It involves measuring the proportion of graduates (excluding those opting to pursue further study or unavailable to work) in full or part time employment within 12 months of graduation.

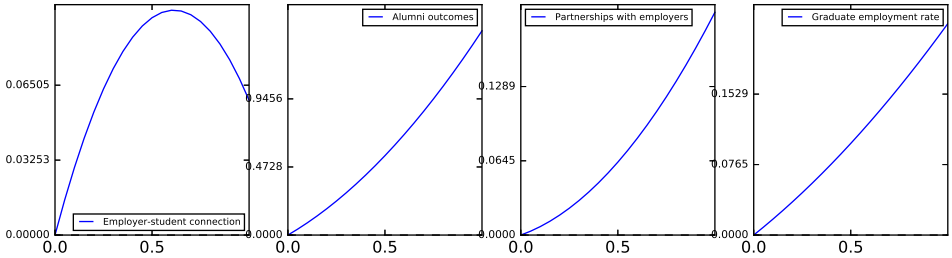

Figure 15 Marginal value functions obtained by NN-MCDA.

Note. We use polynomial functions of degree 2 to approximate the model. The black dashed line is the baseline rate satisfying $p(\hat{y}=1|\mathbf{x})=p(\hat{y}=0|\mathbf{x})=0.5$. Same in following Figures.

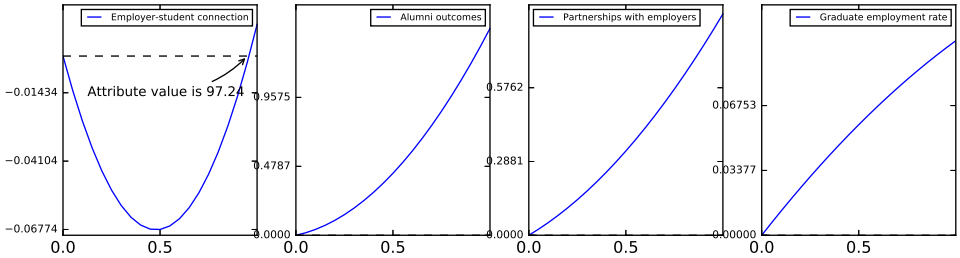

Figure 16 Marginal value functions obtained by logistic regression model.

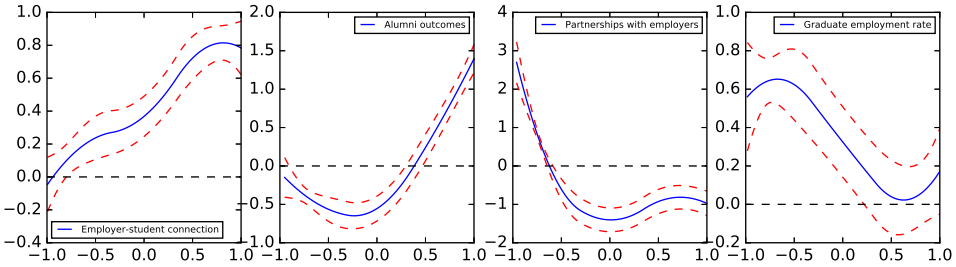

Figure 17 Marginal value functions obtained by GAM with 5 splines that are in 2-degree polynomials.

Note. The red dashed lines are the 95% confidence bands.

common knowledge. For attribute *employer-student connection*, GAM obtains a generally increasing curve. Although it also captures a slight dip of the increasing trend for *employer-student connection*greater than 0.8, such effect is not as clear as that in the curve captured by NN-MCDA. For attributes *alumni outcomes*, *partnerships with employers* and *graduate employment rate*, GAM obtains quite different and unstable curves that are difficult to explain.

Table 7 Descriptive statistics of variables used in experiment.

Outcome Variable	Mean	Std Dev	Minimum	Maximum	Type
Center for Epidemiologic Studies Depression (CES-D)	1.52	2.05	0	8	Categ
Predictor					
Age in years (AGE)	67.42	10.99	18	104	Cont
Degree of education (EDU)	12.83	3.44	0	18	Categ
Marital status (MS)	30.13	18.19	0	74.2	Cont
Out-of-pocket expenditure (OPE)	2978.72	7696.11	0	232255.46	Cont
Body mass index (BMI)	28.24	7.14	0	76.6	Cont

Center for Epidemiologic Studies Depression (CES-D) scale is a self-report measure of the frequency of 20 depressive symptoms during the past week (Radloff 1991). It is one of the most popular index assessing the risk for being depressed (Beardslee et al. 2013, Brent et al. 2015, Garber et al. 2009). Age in years at the end of the survey is calculated from the respondent birth date and beginning survey date. The samples vary from 18 to 104 but very few of them are within 18-45 and 93-104 years (Blazer et al. 1991, Mirowsky and Ross 1992). Degree of education is measured by the years of getting education. It is a categorical variable varying from 0 to 18. Note that for respondents whose degree of education is higher than 17 years, we set the degree of education as 18 (Ladin 2008, Murrell et al. 1983). Marital status is represented by the length of the longest marriage that respondent ever had. It is a continuous variable and varies from 0 to 74.2 years (Pearlin and Johnson 1977, Kessler and Essex 1982, Penninx et al. 1998). Out-of-pocket expenditure refers to the expenses that the respondent pays directly to the health care provider without a third-party (insurer, or State). We consider the out-of-pocket medical expenditure in previous 2 years (Gadit 2004). Body mass index is the weight divided by the square height Bugliari et al. (2016), which determines whether a respondent is overweight (Dong et al. 2004, Luppino et al. 2010, Ross 1994). For more detailed statistical description, see supplementary materials.

4.3. Predicting geriatric depression risk.

Depression is a major cause of emotional suffering in later life. Geriatric depression reduces the quality of older adults' life and increases the risk for acquiring other diseases and committing suicide (Alexopoulos 2005). The existing literature empirically studied the risk factors of geriatric depression but few of them gave insight into how these risk factors affect the prevalence of geriatric depression in details (i.e.: the shape of the marginal value functions). It is more managerially helpful for clinical decision making to prevent older adults from being depressed if we can understand how each risk factor influences the risk for depression at different scales.

The Health and Retirement Study (HRS) is a nationally representative longitudinal study of US adults aged 51 years and older (Bugliari et al. 2016). It has been widely used in many medical studies because of its massive information about older adults demographics, health status, health care utilization and costs, and other useful variables (Pool et al. 2018). We sample the data in 2014 ($N = 17,696$). In this experiment, given five pre-determined attributes (risk factors) that have been found to be associated with geriatric depression, we want to capture the detailed effect of these risk factors at different scales, which are represented by marginal value functions. The descriptive statistics are described in Table 7.

The respondents with CES-D scores higher than or equal to 1 are assumed to be at risk for depression (positive samples). There are totally 9,816 positive samples and 7,880

Table 8 **Averaged AUC and α for training and testing sets.**

	Training set	Testing set	α
NN-MCDA	0.678	0.669	0.393
Logistic Regression	0.621	0.593	1
MLP	0.698	0.668	0
GAM (5 splines)	0.628	0.613	-

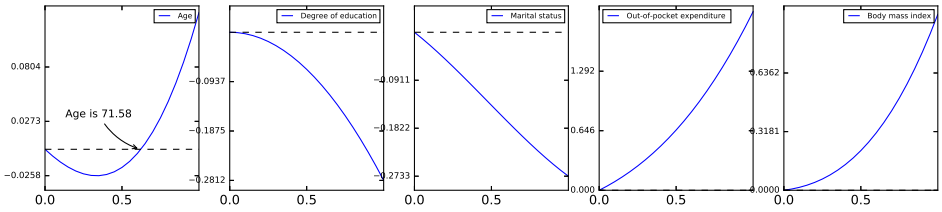

Figure 18 **Marginal value functions obtained by NN-MCDA.**

Note. From the left to right, the attributes are age, degree of education, marital status, out-of-pocket expenditure, and BMI. We use polynomial functions of degree 3 to approximate the model. The black dashed line is the baseline rate satisfying $p(\hat{y}=1|\mathbf{x})=p(\hat{y}=0|\mathbf{x})=0.5$. Same in following Figures.

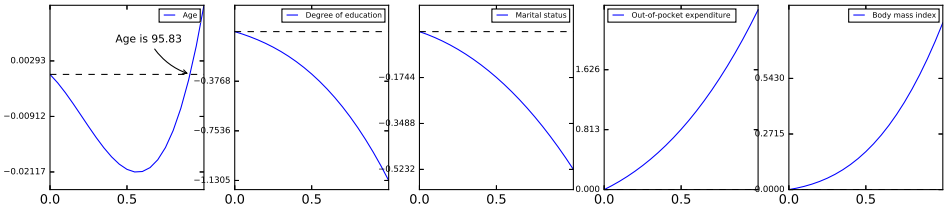

Figure 19 **Marginal value functions obtained by logistic regression model.**

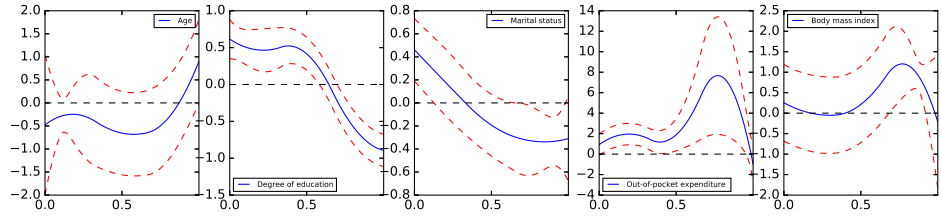

Figure 20 **Marginal value functions obtained by GAM with 5 splines that are in 3-degree polynomials.**

negative samples. We randomly choose 90% from both positive and negative samples to train the model. We train the NN-MCDA, MLP, GAM, and logistic regression model for 30 times, and present the average results in Table 8. The obtained marginal value functions are visualized in Figures 18, 19, and 20.

We first analyze the similar conclusions by three interpretable models. For the last four attributes, both NN-MCDA and logistic regression models capture similar monotonic trends. As for *degree of education* and *marital status*, the curves are under the baseline

rate indicating higher education and the longer length of marriage can reduce the risk for depression. This is consistent with the medical literature (Penninx et al. 1998, Ladin 2008). More specifically, both models find that the attributes *out-of-pocket expenditure* and *Body mass index* would increase the risk for depression. Since the obtained value functions are in a convex shape, the growth of the risk will increase along with the increase of these attribute values.

Both the NN-MCDA and logistic regression obtain a convex curve for attribute *age* with part of the curve being negative and the rest being positive (see Figures 18 and 19). This indicates that the risk of depression does not increase while aging if the adult is younger than a threshold. The risk of depression increases fast after an adult passes an age threshold. The threshold for NN-MCDA is 71.58, which makes sense because most adults younger than 71.58 could be enjoying their retirement and their body functions do not degrade much. However, the threshold for the logistic regression is 95.83, which seems unrealistic and inconsistent with the literature (Blazer et al. 1991).

Similar to the previous experiment, GAM obtains less stabler curves, which are relatively more difficult to interpret. For attribute *age*, GAM obtains similar patterns as NN-MCDA and logistic regression models with an age threshold around 90, which is inconsistent with the literature (Blazer et al. 1991). For attributes *degree of education* and *marital status*, GAM even obtains quite counter-intuitive (if not wrong) results, indicating that the increase in educational and marriage time results in the higher risk of depression.

The experiments with real data demonstrate that the proposed NN-MCDA can effectively capture the patterns in human decision behavior through learning the marginal value functions, which characterize the contribution of individual attributes to the predictions. The NN-MCDA presents good potential in enhancing the empirical studies through providing a detailed marginal value function instead of a single coefficient for each attribute. Moreover, the prediction performance of NN-MCDA is close to a full complexity model (MLP), and much better than that of baseline interpretable models (GAM and logistic regression model).

5. Discussion

In this section, we summarize the insights from the experiments, discuss the use and extension of the proposed NN-MCDA model, and compare the proposed NN-MCDA with ensemble learning.

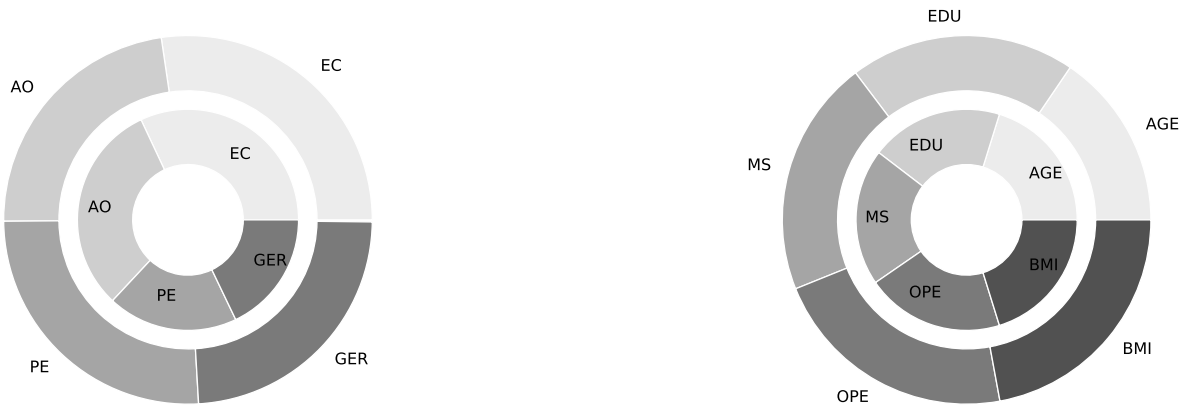

Figure 21 Normalized attribute weights in univerisity employee reputation ranking and geriatric depression prediction.

Note. The inner circles are normalized attribute weights that are obtained by linear regression model and the outer ones are obtained by NN-MCDA.

5.1. Attribute importance.

To explain the importance of each attribute, we present the normalized attribute weights obtained from previous experiments in Figure 21. In the university ranking problem, NN-MCDA assigns 0.2732, 0.228, 0.258, and 0.239 to attributes *employer-student connection*, *alumni outcomes*, *partnerships with employers* and *graduate employment rate* whereas a regression model assigns 0.3198, 0.3107, 0.1906, and 0.1789 to them, respectively. Both models determine that *employer-student connection* is the most important attribute, however, they give different orders of other three attributes. Given the limited resources and the obtained importances of attributes, maintaining a good employer-student connection with frequent employer presences on campus is the most effective method to achieve good employer reputation of a university.

As for the depression prediction problem, a regression model assigns almost equal importance to the five attributes, which is not intuitive for the DM. On the other hand, the NN-MCDA provides a an order of the attributes according to the importance: $BMI \sim OPE \succ MS \sim EDU \succ AGE$ (0.221, 0.218, 0.208, 0.198, 0.155). Such order suggests that obesity and loads of expenditure are the most important risk factors of becoming depressed. While estimating an old adult's risk for depression, a DM (general physician, specialists

and geriatricians) should prioritize the problems related to the older adult's body weight and economical conditions.

5.2. Interpreting the trade-off coefficient.

As a key coefficient, determining the value of α is important to practical applications. It reflects the influence of high-order interactions and complex nonlinearity of variables on the final decisions. In this regard, the NN-MCDA model can be used to explore the complexity of the learning problem. If the convergence α is very small, it indicates that the data is highly complex and the assumption of preference independence is not valid. The DM should then use latest models that account for attribute interactions, such as a Choquet integral-based model (Aggarwal and Fallah Tehrani 2019) or full complexity machine learning models. On the contrary, if α is close to 1, the DM is recommended to use a simpler model to avoid massive computational time and non-interpretable results.

In the case of prediction for depression (Table 8), the convergence α is around 0.4. It indicates that the involved attributes are possibly interacted in this problem. Some related medical studies also empirically demonstrated some interactions between attributes, for example, older adults that are relatively young with higher degree of education may be involved in more social activities and have a more contented life, which lead to lower risk for depression (Li et al. 2014). Moreover, these older adults with longer marriage will obtain more family support and thus they have lower risk for depression (Pearlin and Johnson 1977). We usually address the pairwise interactions because they are easier to interpret and can be visualized by a heating map (Caruana et al. 2015). Given convergence α and marginal value functions obtained *in the presence of* high-order correlations among attributes, we could develop algorithms to use lower-ordered attribute interactions, e.g.: pairwise and triple-wise interactions, to approximate the higher ones. The framework can then be extended to a two-step procedures, including determining marginal value functions and deducing possible lower-ordered correlations.

5.3. Extending the NN-MCDA framework.

The proposed NN-MCDA presents a general modelling framework, which can be easily extended to enhance the performance and adaptivity for various problems. In this section, we discuss the three extensions, including adding regularizations replacing the model in the nonlinear component, and incorporating attributes in the nonlinear component.

5.3.1. Adding regularizations For some mission-critical cases where the data is complex, the convergence α could be very small. However, the DM still requires certain level of model interpretability to facilitate their decision making. Therefore, we opt for adding a regularization term to prevent the model from being too complicated and non-interpretable. The inclusion of the regularization term also helps prevent the over-fitting problem. For example, we can revise the original MSE as follows $MSE_1 = MSE + (1-\alpha)^2$. The added regularization term $(1-\alpha)^2$ allocates more weight to the linear component at the cost of lower fitting accuracy. We can also change the regularization term to $(2\alpha-1)^2$, which leads to a balanced model that favors a model with equal weights to the linear and nonlinear components. The exact form of the loss function should be selected according to the problem settings.

5.3.2. Replace the model in the nonlinear component. Given different types of datasets, the proposed NN-MCDA model can be modified by replacing the neural networks in the nonlinear component by other network structures. In subsection 4.1.1, NN-MCDA has difficulty in handling extremely complex data ($\mathcal{D}^n_{polynomial-3}$ and $\mathcal{D}^n_{polynomial-15}$). To improve the performance on this data, we can introduce more layers in the MLP or increase the number of neurons in each layer. For image classification problem, we can replace the MLP with a convolutional neural network (CNN), and use the features obtained from CNN as the input for the linear component. To fit time-series or free text data, we can replace the MLP with a recurrent neural network.

In addition, the proposed model can be progressively modified by iteratively interacting with the DM. We provide a user-interactive process to determine the ultimate model. The framework is shown in Figure 22 and explained as follows:

Step 1. We first apply the NN-MCDA model to the management problem. While the model converges, we obtain the value of α.

Step 2. If $\alpha > 0.1$, go to step 3. If $\alpha \leq 0.1$, which indicates that the data are potentially very complex. We opt for a full complexity black-box model to achieve higher accuracy and present the results to the DM. If the DM agrees to use the black-box model, the process is end. Otherwise, we add a regularization term (e.g.: use MSE_1) to the original NN-MCDA, and go back to Step 1.

Step 3. If $\alpha < 0.9$, go to step 4. If $\alpha \geq 0.9$, which indicates that an interpretable model is sufficient to fit the data, we explain the results to the DM. If the DM is satisfied with the

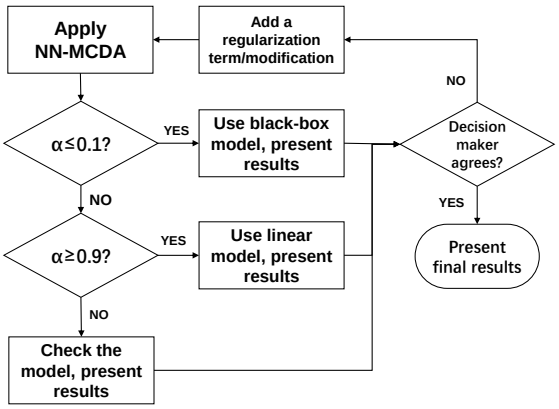

Figure 22 A heuristic user-interactive process for modifying the proposed NN-MCDA.

accuracy, the process ends. Otherwise, we can modify the NN-MCDA model by replacing the model in the nonlinear component (e.g.: using deeper MLP or other neural net-based model). Then, we go back to Step 1.

Step 4. If $0.1 < \alpha < 0.9$, we present the underlying model and results to the DM. If there are no further requirements, the process is end. If the DM requires further modifications (such as adding regularization terms or modifying the nonlinear component), we modify the model accordingly and then go to Step 1.

5.3.3. Flexible inclusion of attributes in the nonlinear component In practice, the human decision behavior usually focuses on a small number of key attributes/criteria (Ribeiro et al. 2016). However, there could exist other minor attributes that do not directly contribute to the prediction, but could affect the prediction through non-traceable complex interactions with other attributes (for example, the interaction between the nonlinear transformation of an attribute and the nonlinear transformation of five other attributes). These minor attributes can be incorporated by the nonlinear component.

In the geriatric depression experiment, the gender of an older adult may not directly indicate a difference in the risk for depression, however, it might still influence the prediction through complex interactions with other attributes. We further extend the NN-MCDA model in incorporate gender and smoking status into the nonlinear component (as shown in Figure 23). We find that the incorporation of these attributes indeed improves the prediction accuracy (the AUC for testing set increases from 0.669 to 0.675), while still maintains the similar marginal value functions in the linear component (see Figure 24). If we add two more attributes, for instance whether the respondent received any home cares in last

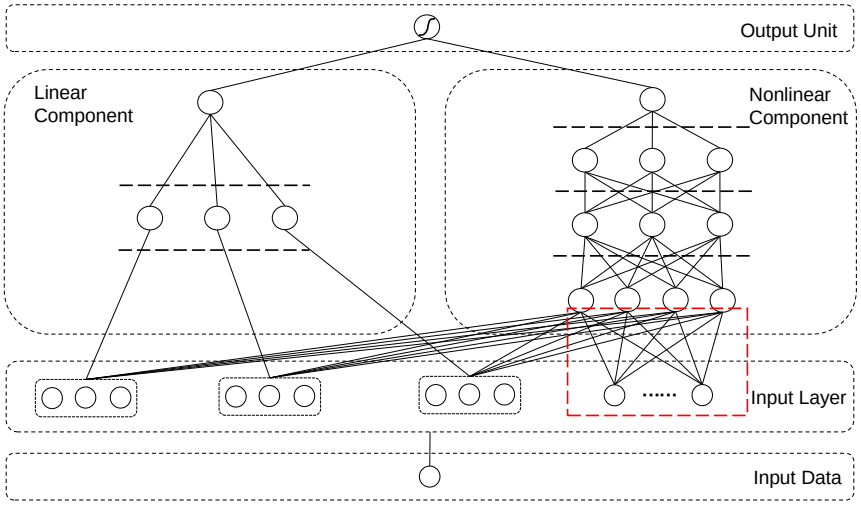

Figure 23 **A new framework for NN-MCDA considering more attributes in the nonlinear component.**

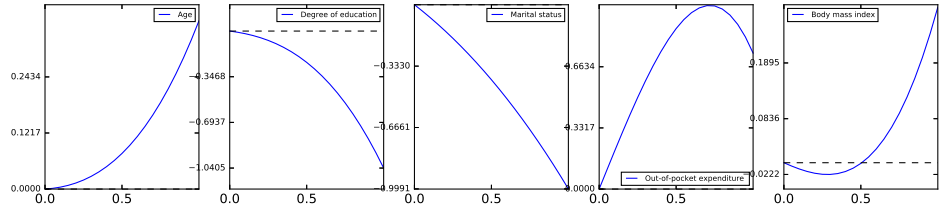

Figure 24 **The marginal value function obtained by linear component while adding two more attributes to the nonlinear component.**

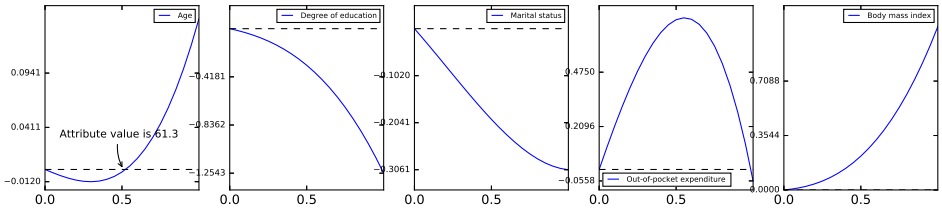

Figure 25 **The marginal value function obtained by linear component while adding four more attributes to the nonlinear component.**

two years and whether the health problem limited his/her work, the AUC increases more obviously (from 0.675 to 0.708) and the marginal value functions still provide convincing results (see Figure 25).

5.4. Joint training process.

In the NN-MCDA model, the linear and nonlinear components are combined by a trade-off coefficient α. Their sum is then fed to a common logistic function for a joint training process. Note that this joint training process is different from ensemble learning (Cheng

et al. 2016), in which multiple classifiers are trained individually and their predictions are simply combined after every model is optimized separately. For example, an ensemble learning approach could have a linear logistic regression model and an MLP model to make predictions for the same dataset separately, and then integrate the prediction results of the two models. The joint training process indicates that the linear and nonlinear components are connected. While we tune the parameters in one component, the other component will be affected. If the model is at its global optimal, the predictions can be made.

6. Conclusion and future work.

In this paper, we proposed a framework for an interpretable model, named NN-MCDA, which combines traditional MCDA model and neural networks. MCDA uses marginal value functions to describe the contribution of individual attributes to the predictions, while neural network considers high-order interrelations among attributes. The framework automatically balances the trade-off between two components. NN-MCDA is more interpretable than a full complexity model and maintains similar predictability.

We present simulation experiments to demonstrate the effectiveness of NN-MCDA. The experiments show that (1) polynomial of higher degrees do not always improve on accuracy; (2) There is a trade-off between the interpretability and the predictability of the model. NN-MCDA can achieve a good balance between them; (3) Given simple data, NN-MCDA performs as good as interpretable model, while given more complex data, NN-MCDA outperforms an interpretable model. We also present how to apply the NN-MCDA framework to real-world decision making problems. These experiments with real data demonstrate the good prediction performance of NN-MCDA and its ability in capturing the detailed contributions of individual attributes.

To the best of our knowledge, this research is the first to introduce the interpretability into machine learning models from the perspective of MCDA. The proposed framework sheds light on how to use MCDA techniques to enhance the interpretability of machine learning models, and how to use machine learning techniques to free MCDA from strong assumptions and enhance its generalizability and predictability.

We envisage the following directions for future researches based on the NN-MCDA framework. First, We can further enhance the interpretability of the model through proposing algorithms to approximate the attribute interactions after obtaining the marginal value

functions. Second, additional simulations are needed to validate the effectiveness of the NN-MCDA variants that introduced in the discussion section. Last but not the least, applying the proposed framework to a variety of real-world decision making and prediction problems constitutes another interesting direction for future work.

References

Aggarwal M, Fallah Tehrani A (2019) Modelling human decision behaviour with preference learning. *INFORMS Journal on Computing* 31(2):318–334, URL `http://dx.doi.org/10.1287/ijoc.2018.0823`.

Alexopoulos GS (2005) Depression in the elderly. *The Lancet* 365(9475):1961–1970.

Angilella S, Corrente S, Greco S, Słowiński R (2014) MUSA-INT: Multicriteria customer satisfaction analysis with interacting criteria. *Omega* 42(1):189–200.

Angilella S, Greco S, Matarazzo B (2010) Non-additive robust ordinal regression: A multiple criteria decision model based on the Choquet integral. *European Journal of Operational Research* 201(1):277–288.

Baesens B, Setiono R, Mues C, Vanthienen J (2003) Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science* 49(3):312–329.

Beardslee WR, Brent DA, Weersing VR, Clarke GN, Porta G, Hollon SD, Gladstone TR, Gallop R, Lynch FL, Iyengar S, et al. (2013) Prevention of depression in at-risk adolescents: longer-term effects. *JAMA Psychiatry* 70(11):1161–1170.

Blazer D, Burchett B, Service C, George LK (1991) The association of age and depression among the elderly: an epidemiologic exploration. *Journal of Gerontology* 46(6):M210–M215.

Brent DA, Brunwasser SM, Hollon SD, Weersing VR, Clarke GN, Dickerson JF, Beardslee WR, Gladstone TR, Porta G, Lynch FL, et al. (2015) Effect of a cognitive-behavioral prevention program on depression 6 years after implementation among at-risk adolescents: a randomized clinical trial. *JAMA Psychiatry* 72(11):1110–1118.

Bugliari D, Campbell N, Chan C, Hayden O, Hurd M, Main R, Mallett J, McCullough C, Meijer E, Moldoff M, et al. (2016) Rand hrs data documentation, version p. *RAND Center for the Study of Aging* .

Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730 (ACM).

Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anderson G, Corrado G, Chai W, Ispir M, et al. (2016) Wide & deep learning for recommender systems. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 7–10 (ACM).

Ciomek K, Ferretti V, Kadziński M (2018) Predictive analytics and disused railways requalification: Insights from a post factum analysis perspective. *Decision Support Systems* 105:34–51.

Corrente S, Greco S, Kadziński M, Słowiński R (2013) Robust ordinal regression in preference learning and ranking. *Machine Learning* 93(2-3):381–422.

Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3):273–297.

Cui D, Curry D (2005) Prediction in marketing using the support vector machine. *Marketing Science* 24(4):595–615.

Cui G, Wong ML, Lui HK (2006) Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science* 52(4):597–612.

Dong C, Sanchez L, Price R (2004) Relationship of obesity to depression: a family-based study. *International Journal of Obesity* 28(6):790.

Doumpos M, Zopounidis C (2011) Preference disaggregation and statistical learning for multicriteria decision support: A review. *European Journal of Operational Research* 209(3):203–214.

Dyer JS, Fishburn PC, Steuer RE, Wallenius J, Zionts S (1992) Multiple criteria decision making, multiattribute utility theory: the next ten years. *Management Science* 38(5):645–654.

Figueira J, Greco S, Ehrgott M (2005) *Multiple criteria decision analysis: state of the art surveys*, volume 78 (Springer Science & Business Media).

Fox J, Glasspool D, Grecu D, Modgil S, South M, Patkar V (2007) Argumentation-based inference and decision making–a medical perspective. *IEEE Intelligent Systems* 22(6):34–41.

Friedman J, Hastie T, Tibshirani R (2001) *The elements of statistical learning* (Springer series in statistics New York).

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 1189–1232.

Gadit AA (2004) Out-of-pocket expenditure for depression among patients attending private community psychiatric clinics in pakistan. *Journal of Mental Health Policy and Economics* 7(1):23–28.

Garber J, Clarke GN, Weersing VR, Beardslee WR, Brent DA, Gladstone TR, DeBar LL, Lynch FL, DAngelo E, Hollon SD, et al. (2009) Prevention of depression in at-risk adolescents: a randomized controlled trial. *JAMA* 301(21):2215–2224.

Gartner D, Kolisch R, Neill DB, Padman R (2015) Machine learning approaches for early DRG classification and resource allocation. *INFORMS Journal on Computing* 27(4):718–734.

Ghaderi M, Ruiz F, Agell N (2017) A linear programming approach for learning non-monotonic additive value functions in multiple criteria decision aiding. *European Journal of Operational Research* 259(3):1073–1084.

Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323.

Greco S, Mousseau V, Słowiński R (2008) Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research* 191(2):416–436.

Gunning D (2017) Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web* .

Hastie T, Tibshirani R (1986) Generalized additive models. *Statistical Science* 1(3):297–318.

Hauser JR (1978) Consumer preference axioms: Behavioral postulates for describing and predicting stochastic choice. *Management Science* 24(13):1331–1341.

Jacquet-Lagreze E, Siskos Y (2001) Preference disaggregation: 20 years of MCDA experience. *European Journal of Operational Research* 130(2):233–245.

Kadziński M, Cinelli M, Ciomek K, Coles SR, Nadagouda MN, Varma RS, Kirwan K (2018) Co-constructive development of a green chemistry-based model for the assessment of nanoparticles synthesis. *European Journal of Operational Research* 264(2):472–490.

Kadziński M, Ghaderi M, Wąsikowski J, Agell N (2017) Expressiveness and robustness measures for the evaluation of an additive value function in multiple criteria preference disaggregation methods: an experimental analysis. *Computers & Operations Research* 87:146–164.

Keeney RL (1976) A group preference axiomatization with cardinal utility. *Management Science* 23(2):140–145.

Keeney RL, Raiffa H (1993) *Decisions with multiple objectives: preferences and value trade-offs* (Cambridge university press).

Kessler RC, Essex M (1982) Marital status and depression: The importance of coping resources. *Social Forces* 61(2):484–507.

Korhonen PJ, Silvennoinen K, Wallenius J, Öörni A (2012) Can a linear value function explain choices? an experimental study. *European Journal of Operational Research* 219(2):360–367.

Ladin K (2008) Risk of late-life depression across 10 european union countries: deconstructing the education effect. *Journal of Aging and Health* 20(6):653–670.

Le QV, Ngiam J, Coates A, Lahiri A, Prochnow B, Ng AY (2011) On optimization methods for deep learning. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 265–272.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436.

Letham B, Rudin C, McCormick TH, Madigan D, et al. (2015) Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9(3):1350–1371.

Li D, Zhang Dj, Shao Jj, Qi Xd, Tian L (2014) A meta-analysis of the prevalence of depressive symptoms in chinese older adults. *Archives of Gerontology and Geriatrics* 58(1):1–9.

Liu J, Liao X, Kadziński M, Słowiński R (2019) Preference disaggregation within the regularization framework for sorting problems with multiple potentially non-monotonic criteria. *European Journal of Operational Research* .

Lou Y, Caruana R, Gehrke J (2012) Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 150–158 (ACM).

Lou Y, Caruana R, Gehrke J, Hooker G (2013) Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 623–631 (ACM).

Luppino FS, de Wit LM, Bouvy PF, Stijnen T, Cuijpers P, Penninx BW, Zitman FG (2010) Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. *Archives of General Psychiatry* 67(3):220–229.

Miller T (2018) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Mirowsky J, Ross CE (1992) Age and depression. *Journal of health and Social Behavior* 187–205.

Mohseni S, Zarei N, Ragan ED (2018) A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839* .

Murrell SA, Himmelfarb S, Wright K (1983) Prevalence of depression and its correlates in older adults. *American Journal of Epidemiology* 117(2):173–185.

Pearlin LI, Johnson JS (1977) Marital status, life-strains and depression. *American Sociological Review* 704–715.

Pelissari R, Oliveira M, Amor SB, Kandakoglu A, Helleno A (2019) Smaa methods and their applications: a literature review and future research directions. *Annals of Operations Research* 1–61.

Penninx BW, Guralnik JM, Ferrucci L, Simonsick EM, Deeg DJ, Wallace RB (1998) Depressive symptoms and physical decline in community-dwelling older persons. *JAMA* 279(21):1720–1726.

Pool LR, Burgard SA, Needham BL, Elliott MR, Langa KM, De Leon CFM (2018) Association of a negative wealth shock with all-cause mortality in middle-aged and older adults in the united states. *JAMA* 319(13):1341–1350.

Radloff LS (1991) The use of the center for epidemiologic studies depression scale in adolescents and young adults. *Journal of Youth and Adolescence* 20(2):149–166.

Ramesh R, Karwan MH, Zionts S (1988) Theory of convex cones in multicriteria decision making. *Annals of Operations Research* 16(1):131–147.

Ramesh R, Karwan MH, Zionts S (1989) Preference structure representation using convex cones in multicriteria integer programming. *Management Science* 35(9):1092–1105.

Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144 (ACM).

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6):386.

Ross CE (1994) Overweight and depression. *Journal of Health and Social Behavior* 63–79.

Rumelhart DE, Hinton GE, Williams RJ (1988) Learning representations by back-propagating errors. *Cognitive Modeling* 5(3):1.

Saaty TL (2013) The modern science of multicriteria decision making and its practical applications: The AHP/ANP approach. *Operations Research* 61(5):1101–1118.

Saaty TL, Decision HTMA (1990) The analytic hierarchy process. *European Journal of Operational Research* 48:9–26.

Sobrie O, Gillis N, Mousseau V, Pirlot M (2018) UTA-poly and UTA-splines: additive value functions with polynomial marginals. *European Journal of Operational Research* 264(2):405–418.

Stewart TJ (1993) Use of piecewise linear value functions in interactive multicriteria decision support: A Monte Carlo study. *Management Science* 39(11):1369–1381.

Vincke P (1986) Analysis of multicriteria decision aid in europe. *European Journal of Operational Research* 25(2):160–168.

Wallenius J, Dyer JS, Fishburn PC, Steuer RE, Zionts S, Deb K (2008) Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management Science* 54(7):1336–1349.

Wang J, Malakooti B (1992) A feedforward neural network for multiple criteria decision making. *Computers & Operations Research* 19(2):151–167.

Zopounidis C, Galariotis E, Doumpos M, Sarri S, Andriosopoulos K (2015) Multiple criteria decision aiding for finance: An updated bibliographic survey. *European Journal of Operational Research* 247(2):339–348.