Failure prediction vs. maintenance prescription: optimizing maintenance interventions by learning individual treatment effects

Toon Vanderschueren[a,d], Robert Boute[b], Tim Verdonck[d], Bart Baesens[a], Wouter Verbeke[b]

[a]*Research Centre for Information Systems Engineering, Faculty of Economics and Business, KU Leuven, 3000 Leuven, Belgium*
toon.vanderschueren@kuleuven.be, bart.baesens@kuleuven.be, wouter.verbeke@kuleuven.be
[b]*Research Centre for Operations Management, Faculty of Economics and Business, KU Leuven, 3000 Leuven, Belgium*
robert.boute@kuleuven.be
[c]*Applied Mathematics, Department of Mathematics, University of Antwerp, 2000 Antwerp, Belgium*
tim.verdonck@kuleuven.be
[d]*Presenting author*

Abstract

Machine maintenance is an important operational problem, where the challenge is to apply sufficient, timely maintenance jobs to avoid machine downtime and maximize its useful life. Data-driven methods can be used to optimize maintenance conditional on the machine's characteristics. Several recent works aim to solve this in the framework of predictive maintenance, where maintenance is planned when the machine's predicted failure probability exceeds a certain threshold. However, this predictive approach does not consider the effect of a maintenance intervention. Therefore, this work proposes a different, prescriptive approach that optimizes maintenance based on the estimated reduction in failure probability resulting from a maintenance intervention. The estimated maintenance effects allow the prescription of the optimal sequence of maintenance interventions and their type during a machine's lifetime. This way, the costs of preventive maintenance and unplanned downtime can be minimized or production output can be maximized. We empirically validate our proposed, prescriptive approach and compare it to a predictive approach using a real-life data set containing detailed information on more than 4,000 full-service maintenance contracts of industrial equipment provided by an industrial partner.

Keywords: Machine maintenance, Prescriptive maintenance, Causal inference, Machine learning

1. Introduction

Machine maintenance is an essential responsibility of asset management that constitutes an important and intricate operational problem. The challenge is to maximize the machine's useful lifetime and avoid expensive downtime due to machine failure, while at the same time preventing unnecessary as well as costly maintenance interventions. Recently, a variety of data-driven solutions have been proposed that consider individual machine characteristics such as historical information on usage and conditions to optimally schedule machine maintenance interventions.

The state-of-the art predictive machine maintenance framework schedules interventions based on risk of failure estimates by a predictive model which makes use of information on the operation

of a machine and machine characteristics (Carvalho et al., 2019). In the predictive framework, a maintenance intervention is planned when the risk of failure exceeds a certain degradation threshold, which balances the cost of treatment and the cost of failure (Poppe et al., 2018).

In this article, we identify a drawback of the predictive approach: *it does not consider the effect of a maintenance intervention on the risk of failure*. This is an important shortcoming that may lead to sub-optimal planning of maintenance interventions. When the predictive model predicts the machine is at risk of failing, a routine maintenance intervention would be scheduled (such as cleaning or oil replenishment). However, at that time, failure might no longer be able to be prevented by a routine maintenance, but might only be remedied by means of more costly repairs or part replacements. Because the predictive framework only considers the risk of failure and not the effect of maintenance, maintenance can be scheduled when it is not at all effective.

This article contributes by proposing a novel prescriptive framework for machine maintenance (Frazzetto et al., 2019; Verbeke et al., 2020; Olaya et al., 2021) that prescribes maintenance based on the estimated effect of a maintenance intervention on the machine's probability of failure. To achieve this, we leverage causal machine learning methods, which can learn models from data to estimate the causal effect of a treatment on an outcome of interest depending on the characteristics of a particular entity, e.g., the causal effect of a maintenance intervention on the risk of failure of an individual machine. Such estimates are called individual treatment effect (ITE) estimates. Moreover, we formulate a prescriptive policy that uses ITE estimates to optimally schedule machine maintenance interventions so as to minimize the total cost of failures and interventions. By assessing the effectiveness of maintenance interventions, our framework allows both to support the development of more effective interventions and to increase the cost-efficiency of maintenance.

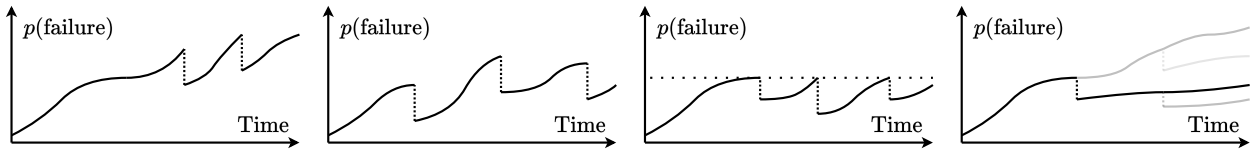
Empirically, we contribute by demonstrating a practical use of the presented prescriptive framework and by evaluating its potential merits using a real-world use case. We present the results of an experimental evaluation of our approach using data on more than 4,000 full-service maintenance contracts of industrial equipment, and compare performance with a state-of-the-art predictive framework.

2. Related work

Machine maintenance has been extensively studied in operations research, with a wide variety of proposed maintenance policies (Wang, 2002; Ding and Kamaruddin, 2015). For a recent overview, we refer the reader to de Jonge and Scarf (2020). In this section, we follow a commonly used categorization of existing work in three general strategies: corrective, preventive and predictive maintenance (Susto et al., 2012, 2014; Carvalho et al., 2019). Finally, we compare these to a new category, prescriptive maintenance, that includes our proposed methodology.

Corrective maintenance. The most elementary approach for machine maintenance is corrective maintenance: in this strategy, maintenance interventions are scheduled as a reaction to failure, with the goal of minimizing the failure's severity (Sheut and Krajewski, 1994). Although this strategy is conceptually simple and prevents unnecessary interventions, it can unavoidably entail significant downtime. Moreover, as corrective maintenance procedures are typically more expensive than preventive interventions, this strategy is typically not cost-efficient.

Preventive maintenance. A more sophisticated, widely studied approach is preventive maintenance (Barlow and Hunter, 1960; Wang, 2002). In this strategy, maintenance interventions are scheduled



(a) Corrective maintenance (b) Preventive maintenance (c) Predictive maintenance (d) Prescriptive maintenance

Figure 1: **Overview of the different maintenance strategies.** (a) In a corrective policy, a maintenance is scheduled as a reaction to a machine failing. (b) Preventive maintenance schedules interventions periodically. (c) A predictive policy monitors the failure probability and schedules interventions when the risk is too high. (d) Prescriptive maintenance estimates the causal effects to consider the different counterfactual scenarios and optimize maintenance interventions.

periodically in time. If failure occurs, maintenance is performed at time of failure and preventive maintenance is rescheduled. This way, the machine's health does not deteriorate drastically and expensive failures are avoided. The difficulty is finding an appropriate time period between maintenance interventions as to avoid both unnecessary interventions and avoidable failures.

Predictive maintenance. As a consequence of recent trends, such as Internet-of-Things and sensor technologies, it has become increasingly easy to measure important machine characteristics between maintenance interventions, leading to an abundance of data at the level of the individual machine. Especially relevant to our work are recent, data-driven approaches that use this data to learn a predictive model (Alaswad and Xiang, 2017). This way, the machine's health can be monitored and data-driven models can predict whether a failure is imminent. When the perceived risk is too high, e.g., exceeding a degradation threshold, an intervention can be scheduled to avoid failure (e.g., Bey-Temsamani et al., 2009; Do et al., 2015; Matyas et al., 2017; Poppe et al., 2018; Nemeth et al., 2018; Ansari et al., 2019).

Prescriptive maintenance. Prescriptive maintenance is a more recent line of work that uses data to prescribe the best maintenance actions. A recent overview is given by Bousdekis et al. (2021). Similar to predictive maintenance, this strategy relies on data-driven decision-making using predictions from a machine learning model. However, in contrast to the predictive approach which relies on the probability of failure to guide decision-making, the focus of these approaches is the maintenance action itself. Existing approaches in this category typically learn a machine-dependent policy by using reinforcement learning (e.g., Rocchetta et al., 2019; Huang et al., 2020; Lepenioti et al., 2020; Ong et al., 2020). Conversely, the prescriptive approach presented in this article makes use of an entirely different class of machine learning models, causal machine learning, to estimate the effect of maintenance intervention. To the best of our knowledge, this approach has not yet been applied in the context of machine maintenance,

3. Methodology

In this section, we introduce a prescriptive framework for optimally scheduling machine maintenance interventions. We start by formalizing the machine maintenance problem in Section 3.1. Subsequently, the state-of-the-art predictive framework, as well as our own prescriptive approach are elaborated and compared in Section 3.2 and Section 3.3. Figure 1 shows an intuitive comparison of our own, prescriptive methodology to the existing maintenance policies detailed in Section 2.

3.1. Problem formulation

We address the problem of scheduling machine maintenance interventions for a pool of machines. To simplify the problem, we discretize continuous time in time slots, s_j for $j \in \{1,\ldots,S\}$. The length of these time slots can be chosen depending on the specific problem domain. A machine i in time slot j is described by a set of characteristics $\mathbf{x}_{i,j} \in \mathbb{R}^d$ for $i \in \{1,\ldots,N\}$ and $j \in \{1,\ldots,S\}$. In each time slot, a binary outcome is observed for each machine, i.e., a failure occurs, $y_{i,j} = 1$, or not, $y_{i,j} = 0$. Failure can generally be defined as unexpected downtime of a machine requiring an intervention to repair it, although the exact definition can depend on the specific problem context.

The challenge for an asset manager is to decide, for each machine and time slot, whether to schedule a maintenance intervention, $t_{i,j} = 1$, or not, $t_{i,j} = 0$. The key goal is to optimally schedule maintenance interventions for each machine individually in order to minimize the total cost resulting from machine failures and maintenance operations, where each failure results in a cost c_f and each maintenance intervention costs c_t.

To facilitate decision-making, the asset manager can leverage a historical data set with information on M machines and maintenance interventions assumed to be drawn from the same distribution: $\mathscr{D} = \{\mathbf{x}_{i,j}, t_{i,j}, y_{i,j}\}_{i=1,j=1}^{i=M,j=S}$. A variety of strategies can be used. This work compares two data-driven approaches: the state-of-the-art predictive framework and our own, prescriptive framework. A high level overview of both approaches is presented in Figure 2.

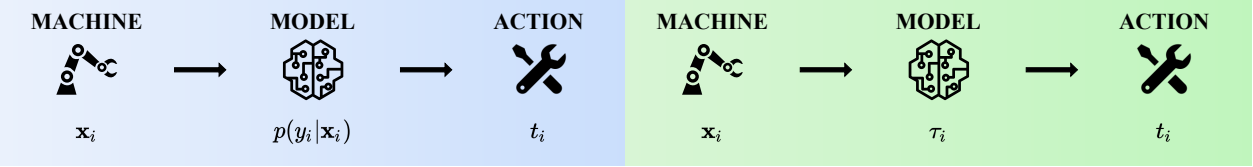
- The predictive machine maintenance framework uses data to learn a *predictive model* for estimating each individual machine's risk of failure in future time slots based on the available information on that machine, $p(y_i|\mathbf{x}_i)$. The predicted failure probability is used as input to a predictive policy for scheduling interventions. Essentially, this framework frames the problem as a cost-sensitive classification task (Elkan, 2001), where machines are classified in the positive or negative class, i.e., to be maintained or not, based on their predicted risk of failure, the cost of a maintenance intervention and the cost of machine failure.

- The proposed prescriptive machine maintenance framework uses data to learn an *ITE model* for estimating the causal effect of a maintenance intervention on the risk of failure for each individual machine in future time slots based on the available information on that machine, $\tau_i = p(y_i|\mathbf{x}_i, t_i = 0) - p(y_i|\mathbf{x}_i, t_i = 1)$. The effect on failure probability is used as input to a prescriptive policy for scheduling interventions. Essentially, this framework frames the problem as a cost-sensitive causal classification task (Verbeke et al., 2020; Olaya et al., 2021), where machines are classified in the positive or negative class, i.e., to be maintained or not, based on the predicted decrease in their risk of failure that would be caused by a maintenance intervention, the cost of a maintenance intervention and the cost of a machine failure.

Conceptually, the prescriptive framework improves upon the predictive framework by not just taking into account the risk of failure, but comparing this risk in two potential scenarios, with and without maintenance, to decide whether or not to schedule a maintenance intervention. Compared to the predictive framework, we conjecture that the prescriptive approach is more aligned with the key objective of machine maintenance: scheduling maintenance to prevent failure where possible. Both frameworks are detailed in the following sections.

3.2. Predictive machine maintenance framework

Scheduling machine maintenance interventions based on failure predictions is identified in Section 3.1 above as a cost-sensitive classification task. Classification is a common machine learning



(a) Predictive maintenance

(b) Prescriptive maintenance

Figure 2: **Comparing predictive and prescriptive maintenance.** Both the predictive (a) and prescriptive approach (b) rely on a data-driven model to leverage machine information $\mathbf{x}_i$ in order to support decision-making, i.e, whether or not to maintain the machine. The key difference is the estimand, which is the conditional failure probability $p(y_i = 1|\mathbf{x}_i)$ for the predictive approach, while the prescriptive approach estimates the conditional effect of a maintenance τ_i.

task that concerns the assignment of machine-instances $\mathbf{x}$ from the instance space $X \subseteq \mathbb{R}^n$ to a class or outcome Y. In this article, machine failure is defined as a binary outcome, yielding a binary classification problem with $Y \in \{0,1\}$.

A binary classifier can learn a binary classification model, $m : X \to [0,1]$, which maps machine-instances x to a positive outcome probability, $P(Y = 1|\mathbf{x})$. A class estimate, $\hat{Y} \in \{0,1\}$, is obtained by setting a classification threshold ϕ. Machines with a positive outcome probability above the threshold, i.e., $P(Y = 1|x) > \phi$, are classified in the positive class, whereas machines with a positive outcome probability below the threshold are classified in the negative class. Machines that are classified in the positive class are expected to fail and hence are to be assigned an intervention in the subsequent time slot.

Hence, the threshold ϕ embeds the predictive decision-making policy to schedule maintenance interventions based on a risk of failure prediction by some classification model. A cost-sensitive threshold, ϕ_{cs}, i.e., a cost-sensitive predictive policy, assigns an intervention to a machine if the cost of an maintenance intervention, c_t, is lower than the *expected cost* of failure, which is calculated based on the cost of failure, c_f and the estimated probability of failure, $\hat{P}(y_i = 1|x_i)$ (Elkan, 2001):

$$\begin{aligned} c_t < c_f \hat{P}(y_i = 1|x_i) &\iff \frac{c_t}{c_f} < \hat{P}(y_i = 1|x_i), \\ &\iff \phi_{cs} < \hat{P}(y_i = 1|x_i). \end{aligned} \tag{1}$$

We find the cost-sensitive threshold for classifying machines in the positive class, i.e., for assigning interventions based on the estimated risk of failure, to be equal to the ratio of the cost of an intervention and the cost of a machine failure:

$$\phi_{cs} = \frac{c_t}{c_f}. \tag{2}$$

Finally, note that well-calibrated probability estimates are required so as to arrive at accurate expected cost estimates and to an optimal predictive policy for scheduling maintenance interventions.

3.3. Prescriptive machine maintenance framework

The proposed prescriptive machine maintenance is introduced in Section 3.1 as a cost-sensitive causal classification task. Causal classification (Fernández and Provost, 2019; Olaya et al., 2021) is defined in terms of the individual treatment effect, τ, which itself is generally defined as the causal effect of a treatment, T, on the outcome, Y, of an instance, $\mathbf{x}_i$. In this study, we employ the

Neyman-Rubin framework to estimate individual treatment effects in terms of potential outcomes, i.e., an individual's outcomes for different treatments (Rubin, 1974). We derive solutions for the double binary case, i.e., with $Y \in \{0,1\}$ and $T \in \{0,1\}$, which can be extended straightforward to multi-treatment settings by performing pairwise comparisons (Haupt and Lessmann, 2022).

In line with the convention for the outcome variable, we refer to treatment $T = 1$ as the positive treatment and to treatment $T = 0$ as the negative treatment. The potential outcome of instance $\mathbf{x}_i$ for treatment $T = t$ is denoted by $Y(\mathbf{x}_i, T = t)$. The individual treatment effect (ITE) is obtained by contrasting the potential outcomes for the positive and the negative treatment.

$$\tau_i := Y(\mathbf{x}_i, T = 1) - Y(\mathbf{x}_i, T = 0). \tag{3}$$

Hence, $\tau_i \in \{-1,0,1\}$, with $\tau_i = 1$ in the context of machine maintenance meaning that a failure is prevented by the maintenance intervention, $\tau_i = 0$ meaning that the intervention did not change the outcome (either failure or no failure), and $\tau_i = -1$ meaning that the maintenance intervention causes a machine failure. With class estimates $P(Y = y|x, T = t)$ and using the expected values for the potential outcomes, $E[Y|x, T = t] = P(Y = 1|x, T = t)$, we obtain ITE estimates, $\hat{\tau} \in [-1,1]$:

$$\hat{\tau}_i := P(Y = 1|\mathbf{x}_i, T = 0) - P(Y = 1|\mathbf{x}_i, T = 1). \tag{4}$$

Note that, by convention, the ITE is defined as the change in the positive outcome probability that is caused by applying the positive treatment, reflecting the objective of minimizing the positive outcome rate among the population, i.e., the number of failures, by means of applying the positive treatment on a subset of the population (Fernández and Provost, 2019). In the context of machine maintenance, the ITE is to be interpreted as the reduction in the risk of machine failure that is caused by the maintenance intervention. Note that, at least theoretically, the ITE can have a negative value, which would mean that the intervention increases the risk of failure.

The objective of causal classification is to identify the machines i with $\tau_i = 1$, i.e., machines where failure can be prevented with maintenance. This way, we minimize both the subset of machine-instances that is to be assigned a maintenance intervention and the number of machine failures, as such optimizing intervention efficiency.

Causal classification is an instance of classification (Fernández and Provost, 2019; Olaya et al., 2021), as formally defined in Section 3.2, where the positive class is the set of instances with $\tau_i = 1$ and the negative class is the set of instances with $\tau_i \in \{-1,0\}$. To causally classify instances in the positive or negative treatment class based on the estimated ITE, i.e., to decide whether to assign an intervention to a machine, a classification threshold ϕ is required. Instances with $\hat{\tau} \geq \phi$ are classified in the positive treatment class and with $\hat{\tau} < \phi$ in the negative treatment class.

A threshold ϕ embeds the prescriptive decision-making policy to schedule maintenance interventions based on ITE estimates. The cost-sensitive threshold, ϕ_{cs}, assigns an intervention to a machine if the expected cost when a maintenance intervention would be carried out is smaller than the expected cost when no maintenance intervention would be carried out. The risk of failure estimates as produced by the ITE model allow to simulate for both scenario's the expected cost:

1. The cost when a maintenance intervention would be carried out: $c_f * p(y_i = 1|x_i, t_i = 1) + c_w$
2. The cost when no maintenance intervention would be carried out: $c_f * p(y_i = 1|x_i, t_i = 0)$

This yields the following cost-sensitive threshold, ϕ_{cs}:

$$\begin{aligned} c_f * p\left(y_i = 1 | x_i, t_i = 1\right) + c_w &< c_f * p\left(y_i = 1 | x_i, t_i = 0\right), \\ \iff \frac{c_t}{c_f} &< p\left(y_i = 1 | x_i, t_i = 0\right) - p\left(y_i | x_i, t_i = 1\right), \\ \iff \frac{c_t}{c_f} &< \hat{\tau}_i. \end{aligned} \tag{5}$$

This threshold can also be obtained by filling in the relevant costs of treatments and benefits of outcomes given our problem formulation in the general formulation of the threshold in the cost-sensitive causal classification framework (Equation 25 in Olaya et al., 2021).

We find the cost-sensitive threshold for causally classifying machines in the positive treatment class, i.e., for assigning interventions based on the estimated ITE, to be equal to the ratio of the cost of an intervention and the cost of a machine failure:

$$\phi_{cs} = \frac{c_t}{c_f}. \tag{6}$$

Note that, even though the cost-sensitive threshold for the prescriptive and predictive policies are identical, the estimand that is used differs. Equation (6) embodies the proposed prescriptive policy based on ITE estimates.

3.4. ITE estimation

Essential to the prescriptive approach is estimating the ITE. This is a well studied problem in fields such as medicine, economics and marketing, which has also been referred to as uplift modeling (Devriendt et al., 2021), heterogeneous treatment effect estimation (Wager and Athey, 2018) and conditional average treatment effect estimation (Shalit et al., 2017). Estimating the ITE differs from standard supervised machine learning as the ITE is never observed in reality. This is because, at each timestep, only a single treatment can be applied to each individual machine and, hence, only one potential outcome, $Y_{t,i}$, is observed for each machine. What would have happened if the machine did not receive a maintenance intervention, the counterfactual scenario, is never observed and, because of this, the ITE itself is never observed. To deal with this, estimating the ITE relies on several standard assumptions: ignorability, common support and stable unit treatment value (SUTVA) (Rubin, 1978; Rosenbaum and Rubin, 1983).

Various machine learning methodologies for learning causal classification models have been proposed in the literature. In general, we can distinguish two main approaches (Olaya et al., 2021). First, metalearners are general strategies to use standard machine learning methods for estimating the ITE. An overview is presented in Künzel et al. (2019). Second, various classification methods have been modified to directly estimate the ITE, such as neural networks (Shalit et al., 2017), causal boosting (Powers et al., 2018) and causal forests (Athey et al., 2019).

In the experiments presented in the following section, we compare the predictive and prescriptive frameworks using four types of machine learning methodologies: logistic regression, support vector machines, random forests and gradient boosting. These classifiers are frequently used for predictive maintenance (Carvalho et al., 2019). For our prescriptive framework, we adapt them to estimate the ITE using a conceptually simple metalearner, the S-learner (Künzel et al., 2019). In this approach, a single machine learning model is trained to predict $\hat{p}(y_i|\mathbf{x}_i, t_i)$ using all data, where only the observed t_i is used during training (i.e., $t_i = 1$ or $t_i = 0$). Then, the ITE can be estimated as $\hat{\tau}_i = \hat{p}(y_i|\mathbf{x}_i, t_i = 0) - \hat{p}(y_i|\mathbf{x}_i, t_i = 1)$.

Type	Attribute	Values
Contract data	Duration in days	$[0, 5843]$
	Contract type	$\{1, 2\}$
	Running hours	$[0, 108508]$
Machine data	ID	/
	Machine type	$\{1, 2, \ldots, 7\}$
	Production date	$\{1979/11/23, \ldots, 2019/07/01\}$
	Industry	36 values
	Country	9 values
	Region	23 values
	Postal code	771 values
	Age at start contract	$\{0, 1, \ldots, 39\}$
	Hours worked at start contract	$[0, 185987]$
Maintenance data	Maintenance interventions	$\{0, 1, \ldots, 33\}$
	Failures over contract period	$\{0, 1, \ldots, 60\}$

Table 1: **Data overview.** We present an overview of the different variables in the data set in terms of what type of data the variable is describing, the name of the attribute, and some information on its possible values.

4. Results

In this section, we present the empirical results of a real-world case study on full-service maintenance contracts where the goal is to optimize maintenance interventions for each piece of equipment. This way, we compare machine maintenance using the two approaches explained in the previous sections: the state-of-the-art predictive approach and our own, prescriptive approach.

4.1. Data

The data consists of more than 4,000 completed full-maintenance contracts for pieces of industrial equipment, which we consider to be single-unit systems. Data is available on the contract itself, the corresponding machine, and the maintenance interventions that were performed during the contract's duration. Table 1 presents an overview of the data.

We simplify the problem setting in several ways. We assume there is only one time slot covering the contract's entire duration. Moreover, we make several adjustments to the data to tackle the problem as a binary classification or binary causal classification task. The number of failures is converted to two condition states, depending on whether equipment i has a relatively low ($y_i = 0$) or high ($y_i = 1$) number of failures per running hours. Similarly, we also constrain the setting to two possible treatments: a relatively low ($t_i = 0$) or high ($t_i = 1$) number of maintenance interventions per running hours. For both, the cutoff point is the median value. In practice, although the exact number of running hours would not be known at the start of the contract when maintenance needs to be planned, an estimate would typically be available.

We assume the contract by default allows for low failure intensity and includes low maintenance intensity without cost. An additional cost c_f is only incurred in case of high failure intensity ($y_i = 1$) and, similarly, a cost c_t in case of a high maintenance intensity ($t_i = 1$), matching the setting described in Section 3.1. Therefore, in the following, we refer to low maintenance intensity as not performing a maintenance intervention and, similarly, to low failure intensity as not failing.

4.2. Evaluation

To evaluate the decisions resulting from the different models and strategies, we need to set an appropriate threshold $\phi_{cs} = \frac{c_t}{c_f}$ (see Equations 2 and 6), for which we would need to know the costs of a maintenance intervention c_t and failure c_f. However, instead of assuming certain costs, we evaluate at different thresholds ϕ_{cs}, corresponding to different cost ratios. Evaluation in this way focuses on the rankings of machines and allows to compare which machines are prioritized for maintenance by the different maintenance strategies.

We empirically evaluate the performance of each model using two types of metrics that quantify two distinct outcomes. The first group looks at the number of failures prevented. The second group evaluates whether failures are correctly predicted. Both are detailed in the following.

Failures prevented. First, we look at the number of failures that are prevented by maintaining machines following the priority given by a model. Evaluating the effect of maintenance is challenging, as the ITE is never observed in reality. To deal with this, we draw inspiration from uplift modeling and look at the realized uplift, which, in this context, corresponds to the number of failures that were prevented due to maintenance. The Uplift@k measures how many failures would have been prevented if the top k% of machines would be prioritized for maintenance. It is calculating by comparing the difference between the average number of observed failures between machines that received a maintenance intervention and those that did not for the first k% ranked machines (Devriendt et al., 2021). Formally, of the first k% ranked instances, consider the number of treated instances $N^k_{\mathscr{T}}$ ($t_i = 1$) of which $\mathscr{T}^k$ failed ($y_i = 1$), and, similarly, let $N^k_{\mathscr{C}}$ be the number of non-maintained machines ($t_i = 0$) of which $\mathscr{C}^k$ failed. Then, the Uplift@k is the difference between the average number of failures in both groups:

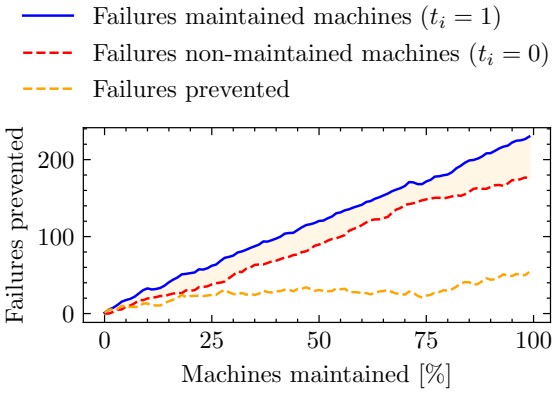


Figure 3: **Uplift curve.** We show an example of an uplift curve that shows the uplift or prevented failures as the difference in failures of maintained and non-maintained machines in the ranking.

$$\text{Uplift@}k = \frac{\mathscr{T}^k}{N^k_{\mathscr{T}}} - \frac{\mathscr{C}^k}{N^k_{\mathscr{C}}}. \tag{7}$$

Similarly, the uplift curve is constructed by looking at the Uplift@k for all possible thresholds up until $k = 100\%$. This curve is summarized by the area under the uplift curve (AUUC), which denotes the area under this curve and compares it to a random (AUUC = 0) and perfect ranking (AUUC = 1) (Devriendt et al., 2020).

Failures predicted. We also evaluate how a model prioritizes machines that are more likely to fail. To do this, we rely on two common evaluation metrics in binary classification: the receiver operating characteristics (ROC) curve and the precision-recall (PR) curve (Davis and Goadrich, 2006). For both, we present a metric that summarizes the curve obtained by a model. For the ROC curve, we use the area under the ROC curve (AUROC). For the PR curve, we use the average precision (AP).

Model	Estimand	Failures prevented				Failures predicted	
		Uplift@10	Uplift@25	Uplift@50	AUUC	AUROC	AP
Logistic regression	Class	0.13	0.11	0.12	-0.06	0.69	0.67
	ITE	0.20	0.16	0.17	-0.02	0.51	0.52
SVM	Class	0.10	0.09	0.12	-0.07	0.70	**0.69**
	ITE	0.16	0.16	0.22	0.03	0.46	0.48
Gradient boosting	Class	0.16	0.14	0.14	-0.05	0.69	0.67
	ITE	0.16	0.17	0.21	0.02	0.49	0.5
Random forest	Class	0.09	0.11	0.12	-0.08	**0.71**	**0.69**
	ITE	**0.42**	**0.37**	**0.29**	**0.15**	0.48	0.49

Table 2: **Results overview**. We compare results when estimating failure probability (class) and maintenance effect (ITE) per model. For each metric, the best result overall is denoted in **bold**, and the best estimand per model is denoted in green. The metrics looking at the failures prevented, uplift@k and area under the uplift curve (AUUC), indicate that models that predict the maintenance effect lead to more prevented failures compared to models predicting failure probability. Conversely, traditional classification metrics, such as the area under the ROC curve (AUROC) or average precision (AP), indicate that classification models are better able to predict failure.

We perform the empirical evaluation using a 10-fold cross-validation, with each fold stratified based on treatment information. We implement all models in `sklearn` and use the default parameters (Pedregosa et al., 2011). This way, the only difference between the classification models and ITE models is the estimand. Finally, categorical variables are transformed using weight-of-evidence encoding (Smith et al., 2002).

4.3. Empirical results

The results of the empirical analysis are shown in Table 2 and the average uplift curve is shown in Figure 4. Models that prioritize machines based on ITE are better at preventing failures, i.e., perform better in terms of uplift@k and AUUC. The models predicting failure probability even prevent less failures than a random policy (AUUC < 0). Conversely, models that prioritize machines based on failure probability are better at predicting failures, i.e. perform better in terms of AUROC and AP. Similarly, most models that estimate ITE perform worse than random in predicting failure (AUROC < 0.5).

The key insight is that models that accurately predict failure do not necessarily result in prevented failures and vice versa, indicating a key trade-off between the two objectives. Importantly, this means that, even though the predictive approach is good at identifying which machines will fail, it does not necessarily result in optimal prevention of failures or, equivalently, optimal maintenance. Conversely, the prescriptive approach does not identify models that are likely to fail, but focuses on the effect of a maintenance intervention to maximize the prevention of failures.

These findings illustrate the importance of focusing machine maintenance on failure prevention (rather than only on detection), which is the key characteristic of our prescriptive approach. To this end, prescribing maintenance for machines with a high risk of failure but with a low estimated ITE points towards ineffective maintenance interventions and a need for other types of interventions.

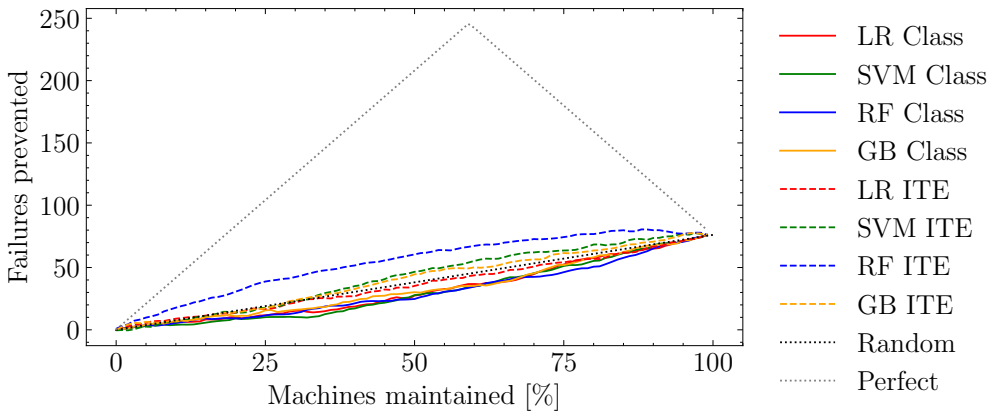


Figure 4: **Failures prevented in terms of machines maintained.** We show the average uplift curves for the different models which indicates how successful a model's prioritization is at preventing failures. Models predicting ITE (in dotted lines) outperform models predicting failure probability (in solid lines). The latter even perform worse than the random model. These findings illustrate the benefit of our prescriptive framework compared to a predictive policy.

These findings imply that the frequently used approach in predictive maintenance to report accuracy metrics (e.g., Susto et al., 2014; Goyal et al., 2016; Su and Huang, 2018; Ayvaz and Alpay, 2021) should not be the only way of measuring performance of a maintenance strategy. The goal is not to accurately predict machine failures but to prevent them, or similarly, to not predict remaining useful life but to maximize it.

5. Conclusion

In this article, we addressed the important operational problem of optimally maintaining machinery. A wide range of existing work tackles this with a predictive approach and plans maintenance when the machine's predicted failure probability exceeds a certain threshold. However, this approach does not take the effect of maintenance interventions into account, which could result in effective maintenance planning. Therefore, this paper contributes by proposing a novel, prescriptive maintenance framework that uses the estimated effects of maintenance interventions based on individual machine characteristics, which is achieved using causal machine learning.

Empirical results indicated major differences between the predictive and prescriptive policies. When compared with our predictive policy, a predictive policy will most urgently plan maintenance interventions for machines with high probability of failing, whereas a prescriptive policy prioritizes maintenance for machines that benefit most from it. Because of this, the prescriptive policy is more effective at failure prevention, which we argue to be the main goal in machine maintenance. These findings do not imply that failure detection is not an essential part of asset management, but rather that it should not be used for maintenance decisions.

The proposed framework opens a range of opportunities for further research. Several degrees of complexity could be added to the problem formulation. For example, depending on the length of the time slots, multiple failures may occur in one time slot and, this way, the outcome of interest would be a discrete number. Alternatively, it could also be relevant to distinguish different types of failures, or similarly, multiple types of maintenance interventions. Moreover, we assumed failure and maintenance costs to be known, machine-independent, static and deterministic, whereas these may be unknown, machine-dependent, dynamic and stochastic. Finally, a limited and potentially

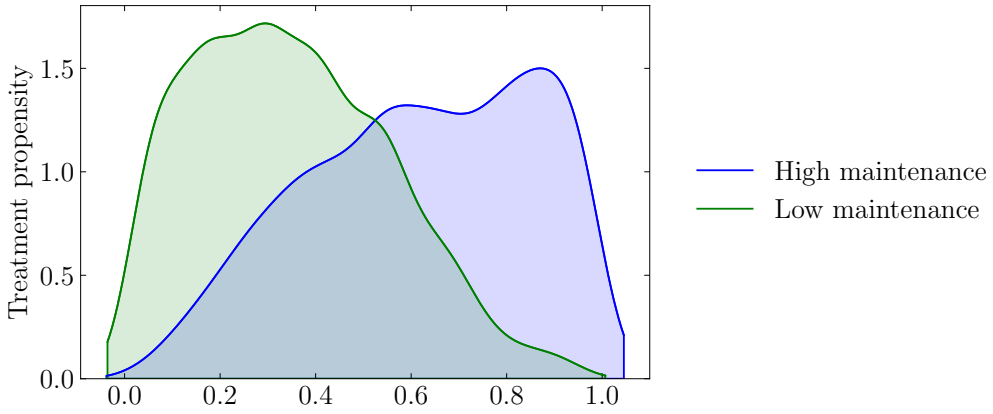


Figure 5: **Selection bias**. The data used in this work is historical, observational data where treatment decisions were made using an existing policy. Because of this, the low and high maintenance groups both have different treatment propensities, i.e., probability to receive the high intensity treatment $p(t_i|\mathbf{x}_i)$. This figure shows the kernel densities of the estimated propensities for both groups using logistic regression with L_2 regularization. This illustrates that treatment assignment was not random, i.e., there is selection bias, as is typically the case with observational data.

stochastic capacity to carry out interventions may impose a constraint on the optimization problem. For all these limitations, extensions to the proposed framework can be conceived to refine the solution to particular requirements or characteristics of other problem settings. One challenge for future research, however, is that public maintenance datasets typically do not contain information on maintenance interventions (Carvalho et al., 2019).

It should also be noted that estimating causal effects is a hard problem. This work did not yet address the important consideration of handling selection bias (see Figure 5). In practice, learning is typically complicated by selection bias when using observational data (Alaa and Schaar, 2018). This is because, historically, not all machines would have been equally likely to receive maintenance, as this would have been prescribed according to an existing policy. Ideally, data would come from a randomized controlled trial where maintenance would be prescribed randomly to different machines, though this is often challenging or excessively expensive in practice. Nevertheless, various methodologies have been proposed that deal with selection bias Hernán and Robins (2006). Furthermore, model validation is challenging in causal machine learning because of the absence of a ground truth Alaa and Van Der Schaar (2019). Finally, causal machine learning methods typically needs to rely on strong assumptions that might be violated in practice, which could cause the causal machine learning methodology to fail (Jesson et al., 2020).

Acknowledgements

This work was supported by the BNP Paribas Fortis Chair in Fraud Analytics, FWO research project G015020N, and FWO PhD Fellowship 11I7322N.

References

Alaa, A., Schaar, M., 2018. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design, in: International Conference on Machine Learning, PMLR. pp. 129–138.

Alaa, A., Van Der Schaar, M., 2019. Validating causal inference models via influence functions, in: International Conference on Machine Learning, PMLR. pp. 191–201.

Alaswad, S., Xiang, Y., 2017. A review on condition-based maintenance optimization models for stochastically deteriorating system. Reliability engineering & system safety 157, 54–63.

Ansari, F., Glawar, R., Nemeth, T., 2019. Prima: a prescriptive maintenance model for cyber-physical production systems. International Journal of Computer Integrated Manufacturing 32, 482–503.

Athey, S., Tibshirani, J., Wager, S., et al., 2019. Generalized random forests. The Annals of Statistics 47, 1148–1178.

Ayvaz, S., Alpay, K., 2021. Predictive maintenance system for production lines in manufacturing: A machine learning approach using iot data in real-time. Expert Systems with Applications 173, 114598.

Barlow, R., Hunter, L., 1960. Optimum preventive maintenance policies. Operations research 8, 90–100.

Bey-Temsamani, A., Engels, M., Motten, A., Vandenplas, S., Ompusunggu, A.P., 2009. A practical approach to combine data mining and prognostics for improved predictive maintenance. Data Min. Case Stud 36.

Bousdekis, A., Lepenioti, K., Apostolou, D., Mentzas, G., 2021. A review of data-driven decision-making methods for industry 4.0 maintenance applications. Electronics 10, 828.

Carvalho, T.P., Soares, F.A., Vita, R., Francisco, R.d.P., Basto, J.P., Alcalá, S.G., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering 137, 106024.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and roc curves, in: Proceedings of the 23rd international conference on Machine learning, pp. 233–240.

Devriendt, F., Berrevoets, J., Verbeke, W., 2021. Why you should stop predicting customer churn and start using uplift models. Information Sciences 548, 497–515.

Devriendt, F., Guns, T., Verbeke, W., 2020. Learning to rank for uplift modeling. IEEE Transactions on Knowledge and Data Engineering doi: 10.1109/TKDE.2020.3048510.

Ding, S.H., Kamaruddin, S., 2015. Maintenance policy optimization—literature review and directions. The international journal of advanced manufacturing technology 76, 1263–1283.

Do, P., Voisin, A., Levrat, E., Iung, B., 2015. A proactive condition-based maintenance strategy with both perfect and imperfect maintenance actions. Reliability Engineering & System Safety 133, 22–32.

Elkan, C., 2001. The foundations of cost-sensitive learning, in: International joint conference on artificial intelligence, Lawrence Erlbaum Associates Ltd. pp. 973–978.

Fernández, C., Provost, F., 2019. Causal classification: Treatment effect vs. outcome prediction. Outcome Prediction .

Frazzetto, D., Nielsen, T.D., Pedersen, T.B., Šikšnys, L., 2019. Prescriptive analytics: a survey of emerging trends and technologies. The VLDB Journal 28, 575–595.

Goyal, A., Aprilia, E., Janssen, G., Kim, Y., Kumar, T., Mueller, R., Phan, D., Raman, A., Schuddebeurs, J., Xiong, J., et al., 2016. Asset health management using predictive and prescriptive analytics for the electric power grid. IBM Journal of Research and Development 60, 4–1.

Haupt, J., Lessmann, S., 2022. Targeting customers under response-dependent costs. European Journal of Operational Research 297, 369–379.

Hernán, M.A., Robins, J.M., 2006. Estimating causal effects from epidemiological data. Journal of Epidemiology & Community Health 60, 578–586.

Huang, J., Chang, Q., Arinez, J., 2020. Deep reinforcement learning based preventive maintenance policy for serial production lines. Expert Systems with Applications 160, 113701.

Jesson, A., Mindermann, S., Shalit, U., Gal, Y., 2020. Identifying causal-effect inference failure with uncertainty-aware models. Advances in Neural Information Processing Systems 33.

de Jonge, B., Scarf, P.A., 2020. A review on maintenance optimization. European journal of operational research 285, 805–824.

Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B., 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences 116, 4156–4165.

Lepenioti, K., Pertselakis, M., Bousdekis, A., Louca, A., Lampathaki, F., Apostolou, D., Mentzas, G., Anastasiou, S., 2020. Machine learning for predictive and prescriptive analytics of operational data in smart manufacturing, in: International Conference on Advanced Information Systems Engineering, Springer. pp. 5–16.

Matyas, K., Nemeth, T., Kovacs, K., Glawar, R., 2017. A procedural approach for realizing prescriptive maintenance planning in manufacturing industries. CIRP Annals 66, 461–464.

Nemeth, T., Ansari, F., Sihn, W., Haslhofer, B., Schindler, A., 2018. Prima-x: A reference model for realizing prescriptive maintenance and assessing its maturity enhanced by machine learning. Procedia CIRP 72, 1039–1044.

Olaya, D., Verbeke, W., Van Belle, J., Guerry, M.A., 2021. To do or not to do: cost-sensitive causal decision-making. arXiv preprint arXiv:2101.01407 .

Ong, K.S.H., Niyato, D., Yuen, C., 2020. Predictive maintenance for edge-based sensor networks: A deep reinforcement learning approach, in: 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), IEEE. pp. 1–6.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Poppe, J., Boute, R.N., Lambrecht, M.R., 2018. A hybrid condition-based maintenance policy for continuously monitored components with two degradation thresholds. European Journal of Operational Research 268, 515–532.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N.H., Hastie, T., Tibshirani, R., 2018. Some methods for heterogeneous treatment effect estimation in high dimensions. Statistics in Medicine 37, 1767–1787.

Rocchetta, R., Bellani, L., Compare, M., Zio, E., Patelli, E., 2019. A reinforcement learning framework for optimal operation and maintenance of power grids. Applied energy 241, 291–301.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688.

Rubin, D.B., 1978. Bayesian inference for causal effects: The role of randomization. The Annals of statistics , 34–58.

Shalit, U., Johansson, F.D., Sontag, D., 2017. Estimating individual treatment effect: generalization bounds and algorithms, in: International Conference on Machine Learning, PMLR. pp. 3076–3085.

Sheut, C., Krajewski, L., 1994. A decision model for corrective maintenance management. The International Journal of Production Research 32, 1365–1382.

Smith, E.P., Lipkovich, I., Ye, K., 2002. Weight-of-evidence (woe): quantitative estimation of probability of impairment for individual and multiple lines of evidence. Human and Ecological Risk Assessment 8, 1585–1596.

Su, C.J., Huang, S.F., 2018. Real-time big data analytics for hard disk drive predictive maintenance. Computers & Electrical Engineering 71, 93–101.

Susto, G.A., Beghi, A., De Luca, C., 2012. A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. IEEE Transactions on Semiconductor Manufacturing 25, 638–649.

Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A., 2014. Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics 11, 812–820.

Verbeke, W., Olaya, D., Berrevoets, J., Verboven, S., Maldonado, S., 2020. The foundations of cost-sensitive causal classification. arXiv preprint arXiv:2007.12582 .

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113, 1228–1242.

Wang, H., 2002. A survey of maintenance policies of deteriorating systems. European journal of operational research 139, 469–489.

Toon Vanderschueren is pursuing a joint PhD in Business Economics at KU Leuven (Belgium) and Mathematics at the University of Antwerp (Belgium). His research focuses on using machine learning to support decision-making in business. He holds bachelor degrees in Business Engineering and in Philosophy, as well as master degrees in Business Engineering and in Artificial Intelligence, all from KU Leuven.

Robert Boute is Full Professor of operations and supply chain management at Vlerick Business School and KU Leuven in Belgium. Robert Boute's research focuses on digital operations, supply chain management and reshoring manufacturing. His recent work includes the use of AI in inventory control and predictive analytics for maintenance optimization.

Tim Verdonck is Professor of Statistics and Data Science at the University of Antwerp (Belgium). He is affiliated to ROBUST@Leuven, the research group on Robust Statistics of KU Leuven. His research interests are in the development and application of data-driven methods for financial, actuarial and economic data sets. He is chairholder of the BNP Paribas Fortis Chair in Fraud Analytics (co-chairholders: Bart Baesens and Wouter Verbeke), the Allianz Chair on Prescriptive Business Analytics in Insurance (co-chairholders: Bart Baesens and Wilfried Lemahieu) and the BASF Chair on Robust Predictive Analytics (co-chairholder Peter Rousseeuw).

Bart Baesens is a professor of Big Data & Analytics at KU Leuven (Belgium), and at the University of Southampton (United Kingdom). He has done extensive research on big data & analytics, credit risk modeling, fraud detection, and marketing analytics. Bart received the OR Society's Goodeve medal for best JORS paper in 2016 and the EURO 2014 and EURO 2017 award for best EJOR paper. His research is summarized at www.dataminingapps.com. Bart is listed in Stanford University's new Database of Top Scientists in the World. He was also named one of the World's top educators in Data Science by CDO magazine in 2021.

Wouter Verbeke is Associate Professor of Data Science at the Faculty of Economics and Business, KU Leuven, Belgium. His research is situated in the field of cost-sensitive and causal machine learning for business decision-making. In 2014, he won the distinguished EURO award for best article published in the European Journal of Operational Research in the category 'Innovative Applications of O.R.' His work has been published in established international scientific journals such as IEEE Transactions on Knowledge and Data Engineering, Information Sciences and European Journal of Operational Research. He has authored two practitioner-oriented books, entitled 'Fraud Analytics Using Descriptive, Predictive & Social Network Techniques' and 'Profit-driven Business Analytics', published by Wiley.