mathematics

MDPI

Article

Solving Contextual Stochastic Optimization Problems through Contextual Distribution Estimation

Xuecheng Tian [1], Bo Jiang [2,*], King-Wah Pang [1], Yu Guo [1], Yong Jin [1] and Shuaian Wang [1]

[1] Faculty of Business, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China; xuecheng-simon.tian@connect.polyu.hk (X.T.); anthony.pang@polyu.edu.hk (K.-W.P.); christine-yu.guo@connect.polyu.hk (Y.G.); jimmy.jin@polyu.edu.hk (Y.J.); hans.wang@polyu.edu.hk (S.W.)

[2] Institute of Data and Information, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

* Correspondence: jb22@mails.tsinghua.edu.cn

Abstract: Stochastic optimization models always assume known probability distributions about uncertain parameters. However, it is unrealistic to know the true distributions. In the era of big data, with the knowledge of informative features related to uncertain parameters, this study aims to estimate the conditional distributions of uncertain parameters directly and solve the resulting contextual stochastic optimization problem by using a set of realizations drawn from estimated distributions, which is called the contextual distribution estimation method. We use an energy scheduling problem as the case study and conduct numerical experiments with real-world data. The results demonstrate that the proposed contextual distribution estimation method offers specific benefits in particular scenarios, resulting in improved decisions. This study contributes to the literature on contextual stochastic optimization problems by introducing the contextual distribution estimation method, which holds practical significance for addressing data-driven uncertain decision problems.

Keywords: data-driven decision making; prescriptive analytics; contextual stochastic optimization

MSC: 90-10

Citation: Tian, X.; Jiang, B.; Pang, K.-W.; Guo, Y.; Jin, Y.; Wang, S. Solving Contextual Stochastic Optimization Problems through Contextual Distribution Estimation. *Mathematics* **2024**, *12*, 1612. https://doi.org/10.3390/math12111612

Academic Editors: Antonio Di Crescenzo and Elvira Di Nardo

Received: 23 March 2024
Revised: 15 May 2024
Accepted: 20 May 2024
Published: 21 May 2024

1. Introduction

Uncertainty is prevalent in various decision-making systems. For optimization problems with uncertain parameters in the objective function, traditional models typically assume that these uncertain parameters follow known probability distributions, thereby establishing stochastic optimization models to solve the problems [1]. However, obtaining the true distributions of uncertain parameters is often challenging. Abundant historical data on uncertain parameters and their related features present new perspectives for addressing uncertainty [2], in fields like transportation [3,4] and logistics [5–7].

Stochastic problems, without knowing feature information, can use a sample average approximation (SAA) to generate approximate stochastic models [8]. However, in the era of big data, we have access to informative features related to uncertain parameters, leading to contextual stochastic problems. When the objective function is linear in uncertain parameters, we can use predict-then-optimize and smart predict-then-optimize methods to predict the point values of uncertain parameters [9]. Then, these point predictions are plugged into downstream optimization problems to derive solutions. However, when the objective function is nonlinear in uncertain parameters, good predictions do not necessarily mean good decisions for uncertain optimization problems [10]. Various machine learning (ML) methods are used to approximate uncertain parameters as a function of features [2]. To some extent, predictive methods lack the capability to consider the influence of prediction errors on downstream decisions. Replacing a random parameter with its mean is widely recognized as potentially resulting in suboptimal solutions for a stochastic optimization

model [11]. Therefore, we should estimate the conditional distributions of uncertain parameters [12]. Weighted SAA (w-SAA) is an advanced method that uses the empirical distributions of uncertain parameters to derive solutions by solving contextual stochastic problems [13]. A global predictive prescription method based on quantile regression is proposed to predict the distributions of the unknown parameters in the optimization model by Wang and Yan [14]. However, the authors do not estimate the mean and variance of the parameter distribution directly.

From the above literature, we emphasize that the existing studies mainly use empirical distributions to solve contextual stochastic problems, which lack the estimation of the exact underlying distributions of uncertain parameters. Consequently, this study builds on w-SAA but goes a step further; that is, we estimate the conditional distributions of uncertain parameters accurately using ML models and then utilize the estimated distribution to generate a set of estimates, which are then used to obtain approximate solutions to contextual stochastic problems. We use an energy scheduling problem as the case study, build the contextual stochastic optimization model, and conduct numerical experiments with real-world data, as well as four specific ML models, including *k*-nearest-neighbors (*k*NN), classification and regression tree (CART), random forest (RF), and kernel regression (KR). The results demonstrate that the proposed contextual distribution estimation method offers specific benefits in particular scenarios, resulting in improved solutions compared to w-SAA and providing new tools for addressing contextual stochastic optimization problems in practice.

The remainder of this study is organized as follows. Section 2 introduces w-SAA and contextual distribution estimation methods. Section 3 presents the mathematical model for our studied case and the results of numerical experiments. Section 4 concludes this study.

2. Methodology

This section provides an overview of the general form of the contextual stochastic optimization problem, introduces w-SAA and contextual distribution estimation methods, and defines the decision loss as the evaluation metric. In addition, we present the specific steps for implementing the two methods by using four ML models.

2.1. Contextual Stochastic Optimization Problem

Consider: (i) Y is a random variable ($Y \in \mathcal{Y} \subset \mathbb{R}^{d_y}$, where d_y is the dimension of Y), with the underlying true distribution μ_Y and historical observed data $\{y^1, \ldots, y^N\}$; (ii) z ($z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$, where d_z is the dimension of z) represents the decision variable, where $\mathcal{Z}$ is the feasible solution set subject to deterministic constraints; and (iii) $c(z;Y)$ denotes the uncertain cost function. The traditional stochastic optimization problem is defined as follows:

$$v^{\text{stoch}} = \min_{z \in \mathcal{Z}} \mathbb{E}[c(z;Y)],$$
$$z^{\text{stoch}} \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}[c(z;Y)]. \quad (1)$$

If we have historical observed data of feature variables $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ (where d_x is the dimension of X) related to Y, denoted as $\{x^1, \ldots, x^N\}$, where x^i and y^i are observed simultaneously for $i \in \{1, \ldots, N\}$, we can establish a dataset:

$$S_N = \left\{ \left(x^1, y^1\right), \ldots, \left(x^N, y^N\right) : x^i \in \mathcal{X}, y^i \in \mathcal{Y}, i \in \{1, \ldots, N\} \right\}.$$

If we have a new observation $X = x^0$, the contextual stochastic optimization problem is established as follows:

$$v^*\left(x^0\right) = \min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y) | X = x^0 \right],$$

$$z^*\left(x^0\right) \in \mathcal{Z}^*\left(x^0\right) = \arg\min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)|X=x^0\right]. \tag{2}$$

2.2. The w-SAA Method

When $c(z;Y)$ is linear in the random variable Y and we have sufficient data, we can utilize an ML model to predict the expected value of the random variable Y, denoted as $\hat{\mathbb{E}}\left[Y|X=x^0\right] = \hat{y}(x^0)$. By plugging this point estimate into the objective function, Problem (2) can be approximated as:

$$\hat{z}_N^{\text{point}}\left(x^0\right) \in \arg\min_{z\in\mathcal{Z}} c\left[z;\hat{y}\left(x^0\right)\right]. \tag{3}$$

However, when $c(z;Y)$ is nonlinear in the random variable Y, a good prediction may not lead to a good solution. Given this case, the w-SAA method is an advanced alternative. This study focuses on stochastic problems whose objective functions are nonlinear in their uncertain parameters.

The w-SAA method assigns weights to each available data sample and then utilizes SAA for the solution [13]. This method approximates Problem (2) as:

$$\hat{z}_N^{\text{w-SAA}}\left(x^0\right) \in \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} w_{N,\,i}\left(x^0\right) c\left(z;y^i\right), \tag{4}$$

where $w_{N,i}(x^0)$ represents the weight assigned to the data sample (x^i, y^i) based on dataset S_N, observation $X = x^0$, and a specific ML method (such as kNN). It is evident that the w-SAA method directly uses the historical data as an empirical distribution, which does not estimate the mean and the variance of the parameter distribution directly.

2.3. The Contextual Distribution Estimation Method

We now aim to go a step further beyond w-SAA by estimating the mean and the variance of the conditional distribution of the random variable Y given an observation. Specifically, given $X = x^0$, we seek to estimate the conditional mean $\mathbb{E}\left[Y|X=x^0\right]$ and the conditional variance $\mathbb{D}\left[Y|X=x^0\right]$. To elaborate, the estimated conditional mean of Y is equivalent to the point prediction of the ML model, denoted as $\hat{\mathbb{E}}\left[Y|X=x^0\right] = \hat{y}(x^0)$. The prediction error of the model on the training dataset is defined as $\epsilon^i = \hat{\mathbb{E}}\left[Y|X=x^i\right] - y^i$, $i \in \{1,\ldots,N\}$ [15]. Consequently, the estimated conditional variance of Y can be calculated as:

$$\hat{\mathbb{D}}\left[Y|X=x^0\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\epsilon^i - \frac{\sum_{i'=1}^{N}\epsilon^{i'}}{N}\right)^2.$$

Then, we plot the frequency distribution of $\{y^1,\ldots,y^N\}$ to fit the random variable Y with a suitable distribution (e.g., a Gaussian distribution). Consequently, we obtain an estimate of the conditional distribution given $X = x^0$, denoted as $\hat{\mu}_{Y|X=x^0}$. Based on this estimated distribution, we can generate a total of U estimates of Y, represented as $\hat{y}_u(x^0), u \in \{1,2,\ldots,U\}$. Subsequently, Problem (2) can be approximated as:

$$\hat{z}_N^{\text{distr\_esti}}\left(x^0\right) \in \arg\min_{z\in\mathcal{Z}} \frac{1}{U}\sum_{u=1}^{U} c\left[z;\hat{y}_u\left(x^0\right)\right]. \tag{5}$$

Finally, Algorithm 1 shows the procedures of the contextual distribution estimation method.

Algorithm 1. The pseudo-code of the contextual distribution estimation method.

Input: A known dataset $S_N = \left\{\left(x^i, y^i\right), x^i \in \mathcal{X}, y^i \in \mathcal{Y}, i = 1, \ldots, N\right\}$, and a new observation $X = x^0$.

Output: The approximate solution $\hat{z}_N^{\text{distr\_esti}}\left(x^0\right)$ for Problem (2).

Step 1. Plot the frequency distribution of $\{y^1, \ldots, y^N\}$ and determine the approximate distribution type of the random variable Y.

Step 2. Employ machine learning models to obtain the estimated conditional mean of the random variable Y, given $X = x^0$:

$$\hat{\mathbb{E}}\left[Y|X = x^0\right] = \hat{y}\left(x^0\right).$$

Calculate the prediction error of the ML model on the training dataset:

$$\epsilon^i = \hat{\mathbb{E}}\left[Y|X = x^i\right] - y^i, \qquad i \in \{1, \ldots, N\}.$$

Obtain the estimated conditional variance of Y, given $X = x^0$:

$$\hat{\mathbb{D}}\left[Y|X = x^0\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\epsilon^i - \frac{\sum_{i'=1}^{N}\epsilon^{i'}}{N}\right)^2.$$

Step 3. Fit the random variable Y with the approximate distribution type determined in Step 1 and the estimated conditional mean and variance, i.e., $\hat{\mathbb{E}}\left[Y|X = x^0\right]$ and $\hat{\mathbb{D}}\left[Y|X = x^0\right]$ in Step 2, and obtain the estimated conditional distribution $\hat{\mu}_{Y|X=x^0}$, given $X = x^0$.

Step 4. Generate a total of U samples from $\hat{\mu}_{Y|X=x^0}$, represented as $\hat{y}_u(x^0), u \in \{1, 2, \ldots, U\}$.

Step 5. Solve the following model and obtain the approximate solution:

$$\hat{z}_N^{\text{distr\_esti}}\left(x^0\right) \in \arg\min_{z \in \mathcal{Z}} \frac{1}{U}\sum_{u=1}^{U} c\left[z; \hat{y}_u\left(x^0\right)\right].$$

2.4. The Evaluation Metric

To evaluate the effectiveness of the above methods in solving Problem (2), we introduce the decision loss. We define the test dataset T_M as $T_M = \left\{\left(x^{N+1}, y^{N+1}\right), \ldots, \left(x^{N+M}, y^{N+M}\right) : x^j \in \mathcal{X}, y^j \in \mathcal{Y}, j \in \{N+1, \ldots, N+M\}\right\}$, where M denotes the number of test data points. Based on the dataset S_N, we can obtain the decision $\hat{z}_N\left(x^j\right)$ for the observation $X = x^j$ using various methods. The optimal objective function value under complete information is $v^*\left(x^j\right) = \min_{z \in \mathcal{Z}} c\left(z; y^j\right)$, $j \in \{N+1, \ldots, N+M\}$. Therefore, the decision loss is defined as:

$$L_N = \frac{1}{M}\sum_{j=N+1}^{N+M}\left\{c\left[\hat{z}_N\left(x^j\right); y^j\right] - v^*\left(x^j\right)\right\},$$

where $c\left[\hat{z}_N\left(x^j\right); y^j\right]$ denotes the actual cost resulting from decision $\hat{z}_N\left(x^j\right)$ for the observation $X = x^j$, with the difference between $c\left[\hat{z}_N\left(x^j\right); y^j\right]$ and $v^*\left(x^j\right)$ representing the decision loss for the observation $X = x^j$. The decision loss L_N, of a certain method based on S_N, is defined as the average of all decision losses on the test dataset T_M.

2.5. ML Methods

Subsequently, we present the specific steps for implementing w-SAA and contextual distribution estimation methods by utilizing four ML models: *k*NN, CART, RF, and KR.

2.5.1. kNN

In the *k*NN regression model, this study adopts the Euclidean distance as the metric for measuring distances. The Euclidean distance between two data samples $x^i = (x_1^i, x_2^i, \ldots, x_{d_x}^i)$ and $x^j = (x_1^j, x_2^j, \ldots, x_{d_x}^j)$ can be represented as:

$$d_{x^i x^j}^{\text{Euclidean}} = \left\| x^i - x^j \right\| = \sqrt{\sum_{r=1}^{d_x} \left(x_r^i - x_r^j \right)^2}.$$

In a trained *k*NN model based on dataset S_N, when $X = x^j, j \in \{N+1, \ldots, N+M\}$, the predicted value of Y is given by:

$$\hat{y}^{k\text{NN}}\left(x^j\right) = \frac{1}{k} \sum_{i \in \mathbb{N}_k(x^j)} y^i,$$

where $\mathbb{N}_k(x^j) = \left\{ i = 1, \ldots, N : \sum_{l=1}^{N} \mathbb{I}\left[\left\| x^j - x^i \right\| \geq \left\| x^j - x^l \right\| \right] \leq k \right\}$ represents the index set of the k-nearest neighbors of x^j; here, k is the hyperparameter of the *k*NN regression model.

In the w-SAA method, the weights assigned to each training data sample $\left(x^i, y^i\right)$ for x^j are as follows:

$$w_{N,\,i}^{k\text{NN}}\left(x^j\right) = \frac{1}{k} \mathbb{I}\left[i \in \mathbb{N}_k\left(x^j\right) \right], \quad i \in \{1,\, \ldots,\, N\}.$$

2.5.2. CART

During the training process of a CART, we input the training dataset S_N and set various hyperparameters of the tree model, including the maximum depth, $depth_{\max}$, the minimum number of samples for splitting, $split_{\min}$, and the minimum number of samples for leaf nodes, $leaf_{\min}$. By tuning these parameters, we obtain the final CART.

The construction process of a CART is as follows:

1. Input training dataset S_N, hyperparameters $depth_{\max}$, $split_{\min}$, and $leaf_{\min}$.
2. For the training dataset $D = \left\{ \left(x^{(1)},\, y^{(1)}\right),\, \ldots,\, \left(x^{(m)},\, y^{(m)}\right) \right\}$ of the current node:
3. If the number of samples is less than $split_{\min}$:

$$m < split_{\min},$$

4. or if the tree depth is greater than or equal to $depth_{\max}$:

$$tree_{\text{depth}} \geq depth_{\max},$$

5. return a decision subtree and stop recursion at the current node.
6. Otherwise, proceed to step 3.
7. Traverse all feature dimensions and feature values of dataset D, and select the feature dimension $x_r,\ r \in \{1,\, 2,\, \ldots,\, d_x\}$, and value s to split the dataset into two parts:

$$D_1(r,\, s) = \{(x,\, y) | x_r \leq s\}$$

$$D_2(r,\, s) = \{(x,\, y) | x_r > s\},$$

8. which minimizes the sum of variances of the left and right subtrees:

$$\min_{(r,s)} \left\{ \sum_{y^{(i)} \in D_1} \left[y^{(i)} - \overline{y}_{D_1} \right]^2 + \sum_{y^{(j)} \in D_2} \left[y^{(j)} - \overline{y}_{D_2} \right]^2 \right\},$$

9. where $\overline{y}_{D_1} = \frac{\sum_{y^{(i)} \in D_1} y^{(i)}}{|D_1|}$, $\overline{y}_{D_2} = \frac{\sum_{y^{(j)} \in D_2} y^{(j)}}{|D_2|}$.
10. Recursively call steps 2–3 for the two datasets generated in step 3 until the termination condition is met.

In a trained CART, the predicted value of Y given x^j, $j \in \{N+1, \ldots, N+M\}$ is the mean of all y^i, $i \in \{1, \ldots, N\}$ contained in the leaf node assigned to x^j:

$$\hat{y}^{\mathrm{CART}}\left(x^j\right) = \frac{\sum_{\{i \in \{1, \ldots, N\}: R(x^i) = R(x^j)\}} y^i}{\left|\left\{l \in \{1, \ldots, N\} : R\left(x^l\right) = R\left(x^j\right)\right\}\right|},$$

where $R(x) \in \{1, \ldots, N_t\}$ denotes the leaf node corresponding to input x and N_t is the number of leaf nodes in the CART.

In the w-SAA method, the weights assigned to each training data sample $\left(x^i, y^i\right)$ for x^j are as follows:

$$w_{N,i}^{\mathrm{CART}}\left(x^j\right) = \frac{\mathbb{I}\left[R\left(x^i\right) = R\left(x^j\right)\right]}{\left|\left\{l \in \{1, \ldots, N\} : R\left(x^l\right) = R\left(x^j\right)\right\}\right|}, \quad i \in \{1, \ldots, N\}.$$

2.5.3. RF

RF is an algorithm that integrates multiple CARTs based on the concept of ensemble learning, with each CART as its basic unit. The hyperparameters of an RF include the forest size T (i.e., how many trees are constructed), the maximum depth of each tree, $depth_{\max}$, the minimum number of samples for splitting, $split_{\min}$, and the minimum number of samples for leaf nodes, $leaf_{\min}$. We first input the training dataset S_N, then set various hyperparameters, and finally obtain the final model through hyperparameter tuning.

The process of building an RF is as follows:

1. Input training dataset S_N with feature dimension d_x and set the forest size T.
2. For each tree t, N training samples are randomly drawn from S_N with replacement to form the training dataset for that tree.
3. Build each decision tree t using the CART algorithm.
4. For each new observation, obtain the final prediction result by averaging the prediction results of all decision trees considered.

The prediction function for decision tree t corresponding to input x can be represented as:

$$f_t(x) = \sum_{n=1}^{N_t} \hat{y}_{tn} \cdot \mathbb{I}[R(x) = n], \quad t \in \{1, \ldots, T\},$$

where N_t represents the number of leaf nodes in decision tree t, $R(x) \in \{1, \ldots, N_t\}$ denotes the leaf node assigned to input x, and $\hat{y}_{tn}$ represents the predicted value of leaf node n of decision tree t, $n \in \{1, \ldots, N_t\}$. Therefore, the prediction function of RF can be expressed as:

$$f^{\mathrm{RF}}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x).$$

In a trained RF, for input x^j, $j \in \{N+1, \ldots, N+M\}$, the predicted value of Y is:

$$\hat{y}^{\mathrm{RF}}\left(x^j\right) = f^{\mathrm{RF}}\left(x^j\right).$$

In the w-SAA method, the weights assigned to each training data sample $\left(x^i, y^i\right)$ for x^j are as follows:

$$w_{N,i}^{\mathrm{RF}}\left(x^j\right) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\left[R^t\left(x^i\right) = R^t\left(x^j\right)\right]}{\left|\left\{l \in \{1, \ldots, N\} : R^t\left(x^l\right) = R^t\left(x^j\right)\right\}\right|}, \quad i \in \{1, \ldots, N\}.$$

2.5.4. KR

KR, as a method for fitting nonlinear models, essentially utilizes kernel functions as weight functions to establish nonlinear regression models. In this study, a Gaussian kernel function is employed to fit the data, shown as follows:

$$K^i(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{(\|x-x^i\|)^2}{2h^2}}, \quad i \in \{1, \ldots, N\},$$

where x is a new observation, x^i is a historical data sample, $\|x - x^i\|$ is the Euclidean distance (L2 norm) between x and x^i, and h is the bandwidth, serving as a hyperparameter of the KR model.

For the observation x^j, $j \in \{N+1, \ldots, N+M\}$, the weights of each training data sample (x^i, y^i), $i \in \{1, \ldots, N\}$ to x^j are calculated using the kernel function as:

$$w^{\mathrm{KR}}_{N,i}\left(x^j\right) = \frac{K^i\left(x^j\right)}{\sum_{l=1}^{N} K^l\left(x^j\right)}, \quad i \in \{1, \ldots, N\}.$$

Thus, the prediction value of Y at x^j is the weighted sum of all y^i in the training dataset:

$$\hat{y}^{\mathrm{KR}}\left(x^j\right) = \sum_{i=1}^{N} w^{\mathrm{KR}}_{N,i}\left(x^j\right) y^i.$$

3. Case Study

This section begins with an energy scheduling problem and its stochastic optimization model. Subsequently, we present the real-world data used in the numerical experiments and preprocess the data. Finally, we train four ML models and calculate the decision losses for w-SAA and contextual distribution estimation methods.

3.1. Energy Scheduling Problem and Its Mathematical Model

This section presents an energy production scheduling problem, where the future 24-hour energy prices are uncertain, requiring the company to plan corresponding energy production schedules and maximize the profit obtained from energy sales on the future day. The definitions of parameters and variables are shown in Table 1.

Table 1. The definitions of the parameters and variables.

Set:	
$\mathbf{T}$	Planning horizon, $\mathbf{T} = \{1, 2, \ldots, 24\}$
Parameters:	
$\widetilde{y}_t$	Uncertain energy price per unit in period t, $\widetilde{y}_t \in Y, \quad t \in \mathbf{T}$
ξ	Base cost of producing one unit of energy, $\xi = 200$
a	Production capacity (maximum output) per unit time, $a = 20$
Decision variables:	
z_t	Production quantity in period t, $z_t \geq 0, \quad z_t \in z, \quad t \in \mathbf{T}$
c_t	Production cost in period t, $c_t \geq 0, t \in \mathbf{T}$, defined as a piecewise function of the production quantity z_t. When z_t lies in the intervals of [0,5], (5,10], (10,15], (15,20], the cost per unit of production in the corresponding interval is ξ, 1.25ξ, 1.5ξ, 1.75ξ, respectively; that is $c_t(z_t) = \begin{cases} \xi z_t, 0 \leq z_t \leq 5 \\ 5\xi + 1.25\xi(z_t - 5), 5 < z_t \leq 10 \\ 11.25\xi + 1.5\xi(z_t - 10), 10 < z_t \leq 15 \\ 18.75\xi + 1.75\xi(z_t - 15), 15 < z_t \leq 20. \end{cases}$
C	Total cost within the planning horizon, $C \geq 0$

The settings of the parameters in Table 1 are designed based on the real situation of an energy company. For example, for energy-consuming enterprises, the larger the production quantity, the higher the unit production cost; therefore, we design a piecewise cost function with the gradually increasing unit production cost. This experimental setup can reflect the essence of a class of decision-making problems that enterprises face in reality. As the energy price per hour is an uncertain parameter, this study establishes the following stochastic optimization model:

Model A:

$$\min \mathbb{E}[f(z;Y)] = \mathbb{E}\left[C - \sum_{t\in\mathbf{T}} \widetilde{y}_t z_t\right] \tag{6}$$

subject to

$$z_t \le a, \quad t \in \mathbf{T} \tag{7}$$

$$c_t \ge \xi z_t, \quad t \in \mathbf{T} \tag{8}$$

$$c_t \ge 5\xi + 1.25\xi(z_t - 5), \quad t \in \mathbf{T} \tag{9}$$

$$c_t \ge 11.25\xi + 1.5\xi(z_t - 10), \quad t \in \mathbf{T} \tag{10}$$

$$c_t \ge 18.75\xi + 1.75\xi(z_t - 15), \quad t \in \mathbf{T} \tag{11}$$

$$C = \sum_{t\in\mathbf{T}} c_t \tag{12}$$

$$z_t \ge 0,\ c_t \ge 0,\ C \ge 0, \quad t \in \mathbf{T}. \tag{13}$$

Objective function (6) minimizes the expected total negative profits from energy sales on a future day. Constraint (7) represents the production capacity constraint per unit time. Constraints (8)–(11) denote the cost calculation formulas per unit time. Constraint (12) is the formula for total cost calculation. Constraint (13) specifies the variable domain constraints.

Assume that we have V estimated scenarios for energy prices, represented by $\hat{y}_{v,t}$, $v \in \{1,\ldots,V\}, t \in \mathbf{T}$, each with a likelihood of $\frac{1}{V}$. Therefore, the approximation of Model A is shown as follows:

Model B:

$$\min \frac{1}{V}\sum_{v=1}^{V}\left(C - \sum_{t\in\mathbf{T}} \hat{y}_{v,t} z_t\right) \tag{14}$$

subject to Constraints (7)–(13).

3.2. Introduction of Datasets

The energy price dataset used in this study consists of 14,592 records [16]. We divide it into a training dataset S_N ($N = 11640$) and a test dataset T_M ($M = 2952$). Each record contains historical values of a random variable Y and feature variables X. Specifically, the feature vector $X = (x_1, \ldots, x_9) \in \mathcal{X} \subset \mathbb{R}^9$ and the random variable $Y \in \mathcal{Y} \subset \mathbb{R}$, along with their practical meanings and ranges, are detailed in Table 2.

Table 2. Description of the data features.

Notation	Practical Meaning	Range
x_1	month_of_year	$x_1 \in \{1, 2, \ldots, 12\}$
x_2	week_of_year	$x_2 \in \{1, 2, \ldots, 52\}$
x_3	day_of_week	$x_3 \in \{1, 2, \ldots, 7\}$
x_4	hour_of_day	$x_4 \in \{1, 2, \ldots, 24\}$
x_5	holiday_flag	$x_5 \in \{0, 1\}$
x_6	forecast_wind_production	$x_6 \ge 0$
x_7	forecast_system_load	$x_7 \ge 0$
x_8	forecast_system_marginal_price	$x_8 \ge 0$
x_9	CO_2_intensity	$x_9 \ge 0$
Y	fuel_price	$Y \ge 0$

Specifically, $x_1 = 1$ indicates January of the current year, $x_2 = 1$ indicates the first week of the current year, $x_3 = 1$ indicates Monday, $x_4 = 1$ indicates the first hour of the day, and so on. $x_5 = 1$ indicates that the day is a holiday.

Histograms and density curves of the frequency distribution of the random variable Y based on $\{y^1, \ldots, y^N\}$ are plotted, as shown in Figure 1. From the graph, it can be observed that, without considering some extreme values, the distribution of the random variable Y approximates a normal distribution. Therefore, in the contextual distribution estimation method, this study chooses a normal distribution to fit the random variable Y.

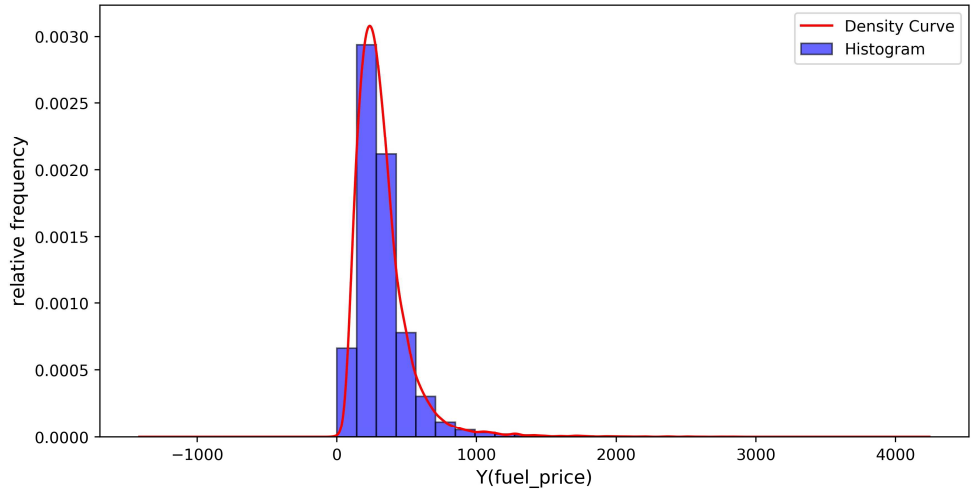

Figure 1. The distribution of the variable Y based on $\{y^1, \ldots, y^N\}$.

Since the mathematical model in this study is aimed at scheduling energy production within a 24-hour period in the future, during testing on the test dataset T_M, each of the 24 data samples form an input of a decision problem. Therefore, we construct a set of decision problems $\{p_w\}$, where $w \in \{1, \ldots, W\}$ and $W = \lfloor M/24 \rfloor = \lfloor 2952/24 \rfloor = 123$.

For problem p_w, $w \in \{1, \ldots, W\}$, its input data is defined as:

$$T(p_w) = \left\{ \left(x^{w1}, y^{w1} \right), \ldots, \left(x^{w24}, y^{w24} \right) \right\},$$

where

$$x^{p_w} = \left[x^{w1}, \ldots, x^{w24} \right],$$

$$y^{p_w} = \left[y^{w1}, \ldots, y^{w24} \right].$$

In the w-SAA method, for x^{wl}, $l \in \{1, \ldots, 24\}$, the weight $w_{N,\,i}\left(x^{wl}\right)$ can be obtained through an ML model, where $i \in \{1, \ldots, N\}$. Therefore, there are N^{24} potential combinations of energy prices in one day. The weight for each parameter combination is $\prod_{l=1}^{24} w_{N,\,i_l}\left(x^{wl}\right)$, where $i_l \in \{1, \ldots, N\}$. This study randomly selects V scenarios along with their weights to be inputted into the objective function of the mathematical model for solving, and obtains the decision $\hat{z}_N^{\mathrm{w-SAA}}(x^{p_w})$. In the contextual distribution estimation method, for x^{wl}, $l \in \{1, \ldots, 24\}$, the estimated distribution $\hat{\mu}_{Y|X=x^{wl}}$ can be obtained. Based on $\hat{\mu}_{Y|X=x^{wl}}$, we obtain U estimates of Y, denoted as $\hat{y}_u\left(x^{wl}\right), u \in \{1, 2, \ldots, U\}$. Therefore, there are U^{24} potential combinations of energy prices in one day, with equal weight for each parameter combination. This study also randomly selects V scenarios to be inputted into the objective function of the mathematical model for solving, and obtains the decision $\hat{z}_N^{\mathrm{distr\_esti}}(x^{p_w})$.

The optimal objective function value of problem p_w under complete information is $v^*(x^{p_w}) = \min_{z \in \mathcal{Z}} f(z; y^{p_w})$; hence, the decision loss is:

$$L_N = \frac{1}{W} \sum_{w=1}^{W} \{ f[\hat{z}_N(x^{p_w}); y^{p_w}] - v^*(x^{p_w}) \}.$$

Based on the training dataset S_N, this study adopts four ML models, i.e., *k*NN, CART, RF, and KR, to compare the effectiveness of the w-SAA and contextual distribution estimation methods.

The *k*NN and KR models involve calculating distances between data samples, thus requiring data standardization before training the models, while the CART and RF models can be trained directly using the original dataset. The standardization method is used as follows: features x_1 to x_4 represent periodic features related to time; therefore, we employ Sine-Cosine encoding, where $x' = \sin\left(2\pi \frac{x}{T_x}\right)$, with x as the original value, x' as the standardized value, and T_x as the total periods of x. For example, the standardized data of the feature $x_1 = (x_1^1, x_1^2, \ldots, x_1^N)$ representing months is $x_1' = (x_1'^1, x_1'^2, \ldots, x_1'^N)$, where:

$$x_1'^i = \sin\left(2\pi \frac{x_1^i}{12}\right), \quad i \in \{1, \ldots, N\}.$$

For features x_5 to x_9, we apply the Z-score standardization method. For instance, for the feature $x_6 = (x_6^1, x_6^2, \ldots, x_6^N)$, the mean is calculated as $\mu_6 = \frac{1}{N}\sum_{i=1}^{N} x_6^i$, and the standard deviation as $\sigma_6 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_6^i - \mu_6)^2}$, resulting in the standardized data $x_6' = (x_6'^1, x_6'^2, \ldots, x_6'^N)$, where:

$$x_6'^i = \frac{x_6^i - \mu_6}{\sigma_6}, \quad i \in \{1, \ldots, N\}.$$

The data after Z-score standardization have a mean of 0 and a standard deviation of 1.

3.3. Model Training

During the model training process, this study randomly splits the training dataset S_N into S_N^{train} and S_N^{valid} in a 4:1 ratio, with S_N^{train} containing 9312 data samples and S_N^{valid} containing 2328 data samples.

In the w-SAA method, the study trains ML models based on S_N^{train} and tunes hyperparameters based on S_N^{valid}, with decision loss as the training metric. In the process of validation and testing, this study randomly selects $V = 200$ scenarios of energy prices to be inputted into the model for solving, yielding approximate solutions $\hat{z}_N^{\text{w-SAA}}$ and calculating the corresponding decision losses $L_N^{\text{w-SAA}}$. The hyperparameter settings and tuning results for the four models are shown in Table 3.

Table 3. Hyperparameter settings and tuning results based on decision loss for w-SAA.

Model	Hyperparameters	Search Range	Optimal Value
*k*NN	*k*	{1, 2, . . ., 50}	36
CART	max_depth	{8, 10, 12, 14}	10
	min_samples_split	{5, 10, 20, 30}	20
	min_samples_leaf	{2, 5, 10, 15}	5
RF	n_estimators	{100, 200}	100
	max_depth	{8, 10, 12}	12
	min_samples_split	{5, 10, 20}	10
	min_samples_leaf	{2, 5, 10}	2
KR	bandwidth	{1, 2, . . ., 20}	4

The line charts of decision loss during the hyperparameter tuning process of the four models for w-SAA are depicted in Figure 2.

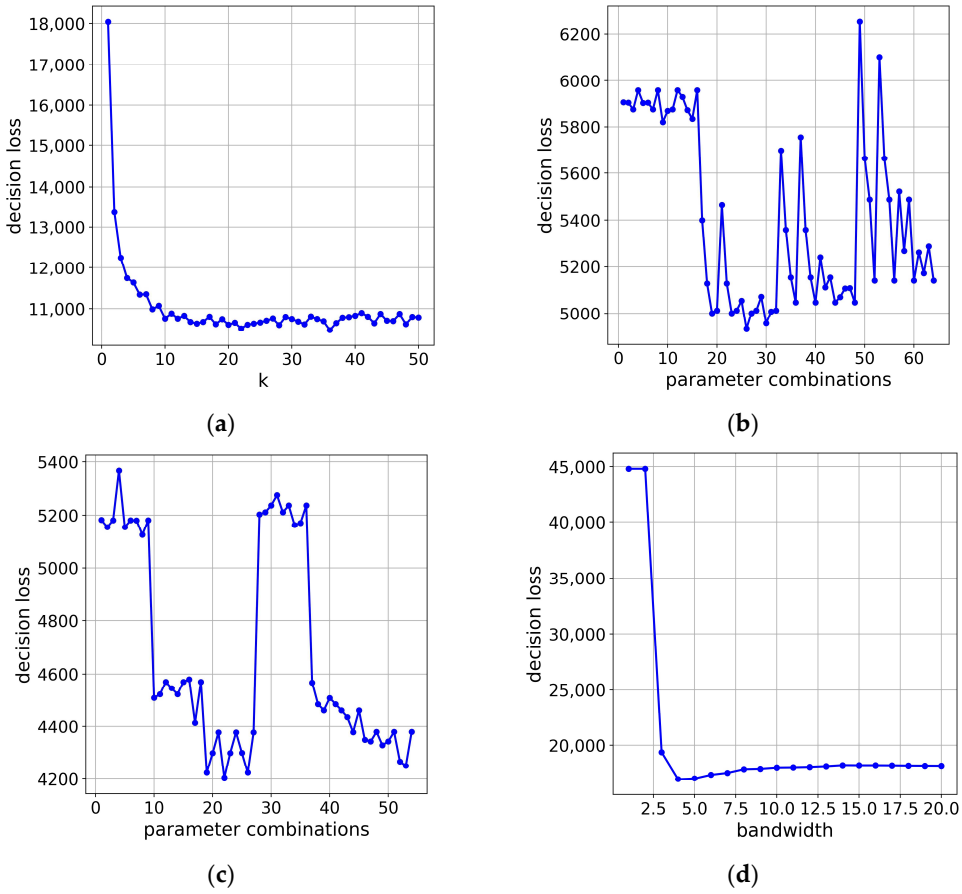

Figure 2. Decision loss on S_N^{valid} of the four models for w-SAA: (**a**) kNN; (**b**) CART; (**c**) RF; and (**d**) KR.

In the contextual distribution estimation method, this study trains ML models based on S_N^{train} and tunes hyperparameters based on S_N^{valid}, with decision loss as the training metric. In the process of validation and testing, this study generates $U = 200$ estimates of energy price based on the estimated distribution $\hat{\mu}_{Y|X=x^{wl}}$, $l \in \{1, \ldots, 24\}$ and randomly selects $V = 200$ scenarios of energy prices to be inputted into the model for solving, yielding approximate solutions $\hat{z}_N^{\text{distr\_esti}}$ and calculating the corresponding decision losses $L_N^{\text{distr\_esti}}$. The hyperparameter settings and tuning results for the four models are shown in Table 4.

Table 4. Hyperparameter settings and tuning results based on decision loss for contextual distribution estimation.

Model	Hyperparameters	Search Range	Optimal Value
kNN	k	{1, 2, . . ., 50}	38
CART	max_depth	{8, 10, 12, 14}	10
	min_samples_split	{5, 10, 20, 30}	30
	min_samples_leaf	{2, 5, 10, 15}	5
RF	n_estimators	{100, 200}	200
	max_depth	{8, 10, 12}	12
	min_samples_split	{5, 10, 20}	5
	min_samples_leaf	{2, 5, 10}	2
KR	bandwidth	{1, 2, . . ., 20}	2

The line charts of decision loss during the hyperparameter tuning process of the four models for contextual distribution estimation are depicted in Figure 3.

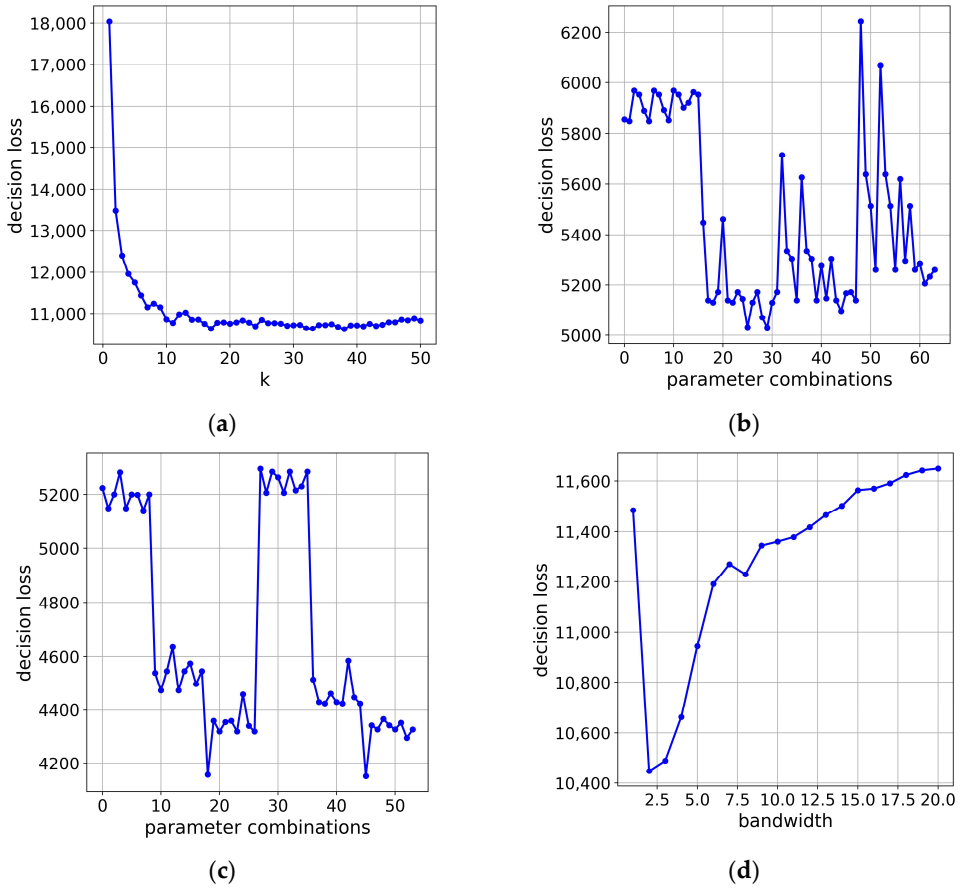

Figure 3. Decision loss on S_N^{valid} of the four models for contextual distribution estimation: (**a**) kNN; (**b**) CART; (**c**) RF; and (**d**) KR.

3.4. Discussion

The experiments are conducted on a computer with AMD Ryzen 5 4600U and 16 GB (3200 MHz) RAM under the Windows 10 operating system. The mathematical model in Section 3.1 is implemented in Python programming language using Gurobi 9.5.2 as the solver. This study trains four ML models, i.e., kNN, CART, RF, and KR, based on the dataset S_N to implement two methods, w-SAA and contextual distribution estimation. The models are tested on the test dataset T_M, and the corresponding decision losses L_N are calculated and summarized in Table 5. Figure 4 illustrates the results.

Table 5. Decision loss of w-SAA and contextual distribution estimation on the test dataset T_M.

	***k*NN**	**CART**	**RF**	**KR**
w-SAA	12,887.14	4753.40	4261.88	18,191.83
distr_esti	13,120.69	4783.65	4408.60	12,668.46

From Table 5 and Figure 4, we can see that compared to the traditional w-SAA method, the proposed contextual distribution estimation method has similar decision loss under the kNN, CART, and RF models, and exhibits certain advantages under the KR models. Specifically, under the kNN model, the decision loss obtained by w-SAA is 1.78% lower than that of our proposed method; under the CART model, the decision loss obtained by w-SAA is 0.63% lower than that of our proposed method; under the RF model, the decision loss obtained by w-SAA is 3.33% lower than that of our proposed method; however, under

the KR model, the decision loss obtained by our proposed method is 30.36% lower than that of w-SAA.

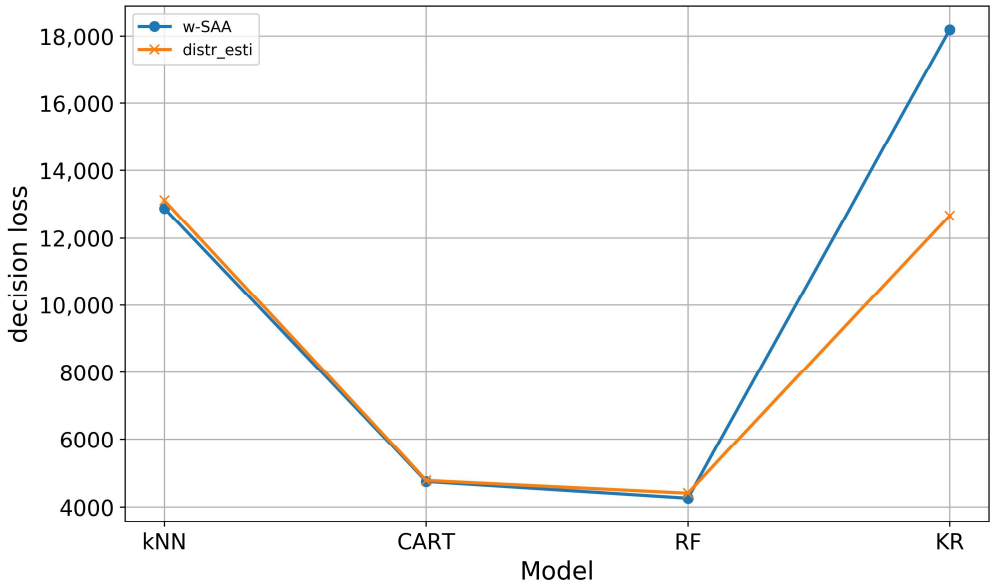

Figure 4. Decision loss of w-SAA and the contextual distribution estimation on the test dataset T_M.

The results of the numerical experiments demonstrate that our proposed method, i.e., contextual distribution estimation, exhibits certain advantages under some particular scenarios, leading to a significant reduction in decision loss.

4. Conclusions

For contextual stochastic optimization problems whose objective functions are nonlinear in their uncertain parameters, this study builds on w-SAA and further estimates the conditional distributions of uncertain parameters accurately, thereby obtaining approximate solutions to the problems. Specifically, we use the point prediction of an ML model as an estimate of the conditional mean, the estimated variance of the differences between predicted values and actual values on the training dataset as estimates of the conditional variance, and fit the uncertain parameters with an appropriate distribution based on historical data. By generating a set of estimates from the estimated distribution and inputting them into the model, an approximate solution to the stochastic optimization problem can be obtained.

This study uses the energy scheduling problem as the case study. Four ML models, i.e., *k*NN, CART, RF, and KR, are trained based on a real-world energy price dataset to implement w-SAA and contextual distribution estimation methods, and the performance of the two methods is tested. From the experimental results, it is shown that the proposed contextual distribution estimation method in this study exhibits advantages in certain scenarios compared to w-SAA, significantly reducing decision losses.

This study introduces the contextual distribution estimation method for contextual stochastic optimization problems, which can be applied to address data-driven uncertain decision problems in the field of operations research and management science, such as transportation and logistics. In future research, extensive computational experiments with data from different fields, such as transportation, manufacturing, and logistics, should be conducted to validate the effectiveness of the contextual distribution estimation method.

Author Contributions: Conceptualization, X.T. and S.W.; methodology, X.T., B.J., K.-W.P., Y.G., Y.J. and S.W.; software, B.J.; validation, B.J., Y.G., Y.J. and K.-W.P.; formal analysis, B.J.; investigation, B.J.; resources, Y.J.; data curation, B.J.; writing—original draft preparation, B.J.; writing—review and editing, X.T. and S.W.; visualization, Y.G.; supervision, S.W.; project administration, K.-W.P.; funding acquisition, Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by AF Competitive Grants of The Hong Kong Polytechnic University (Project ID: P0046074).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and constructive suggestions, which have greatly improved the quality of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Birge, J.R.; Louveaux, F. *Introduction to Stochastic Programming*; Springer Series in Operations Research and Financial Engineering; Springer: New York, NY, USA, 2011; ISBN 978-1-4614-0236-7.
2. Qi, M.; Shen, Z.-J. (Max) Integrating Prediction/Estimation and Optimization with Applications in Operations Management. In *Tutorials in Operations Research: Emerging and Impactful Topics in Operations*; Chou, M., Gibson, H., Staats, B., Shier, D., Greenberg, H.J., Eds.; INFORMS: Catonsville, MD, USA, 2022; pp. 36–58; ISBN 978-0-9906153-7-8.
3. Liu, Y.; Francis, A.; Hollauer, C.; Lawson, M.C.; Shaikh, O.; Cotsman, A.; Bhardwaj, K.; Banboukian, A.; Li, M.; Webb, A.; et al. Reliability of Electric Vehicle Charging Infrastructure: A Cross-lingual Deep Learning Approach. *Commun. Transp. Res.* **2023**, *3*, 100095. [CrossRef]
4. Xu, M.; Di, Y.; Ding, H.; Zhu, Z.; Chen, X.; Yang, H. AGNP: Network-Wide Short-Term Probabilistic Traffic Speed Prediction and Imputation. *Commun. Transp. Res.* **2023**, *3*, 100099. [CrossRef]
5. Qu, X.; Lin, H.; Liu, Y. Envisioning the Future of Transportation: Inspiration of ChatGPT and Large Models. *Commun. Transp. Res.* **2023**, *3*, 100103. [CrossRef]
6. Zhen, L.; Xu, Z.; Wang, K.; Ding, Y. Multi-Period Yard Template Planning in Container Terminals. *Transp. Res. Part B Methodol.* **2016**, *93*, 700–719. [CrossRef]
7. Zhen, L. Modeling of Yard Congestion and Optimization of Yard Template in Container Ports. *Transp. Res. Part B Methodol.* **2016**, *90*, 83–104. [CrossRef]
8. Kleywegt, A.J.; Shapiro, A.; Homem-de-Mello, T. The Sample Average Approximation Method for Stochastic Discrete Optimization. *SIAM J. Optim.* **2002**, *12*, 479–502. [CrossRef]
9. Elmachtoub, A.N.; Grigas, P. Smart "Predict, Then Optimize". *Manag. Sci.* **2022**, *68*, 9–26. [CrossRef]
10. Bertsimas, D.; Koduri, N. Data-Driven Optimization: A Reproducing Kernel Hilbert Space Approach. *Oper. Res.* **2022**, *70*, 454–471. [CrossRef]
11. Shapiro, A.; Dentcheva, D.; Ruszczyński, A. *Lectures on Stochastic Programming: Modeling and Theory*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2021; ISBN 978-0-89871-687-0.
12. Tian, X.; Yan, R.; Wang, S.; Liu, Y.; Zhen, L. Tutorial on Prescriptive Analytics for Logistics: What to Predict and How to Predict. *Electron. Res. Arch.* **2023**, *31*, 2265–2285. [CrossRef]
13. Bertsimas, D.; Kallus, N. From Predictive to Prescriptive Analytics. *Manag. Sci.* **2020**, *66*, 1025–1044. [CrossRef]
14. Wang, S.; Yan, R. "Predict, Then Optimize" with Quantile Regression: A Global Method from Predictive to Prescriptive Analytics and Applications to Multimodal Transportation. *Multimodal Transp.* **2022**, *1*, 100035. [CrossRef]
15. Sadana, U.; Chenreddy, A.; Delage, E.; Forel, A.; Frejinger, E.; Vidal, T. A Survey of Contextual Optimization Methods for Decision-Making under Uncertainty. *Eur. J. Oper. Res.* **2024**, S0377221724002200. [CrossRef]
16. Ifrim, G.; O'Sullivan, B.; Simonis, H. Properties of Energy-Price Forecasts for Scheduling. In *Principles and Practice of Constraint Programming*; Milano, M., Ed.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7514, pp. 957–972; ISBN 978-3-642-33557-0.