mathematics

MDPI

Article

Data Transformation in the Predict-Then-Optimize Framework: Enhancing Decision Making under Uncertainty

Xuecheng Tian, Yanxia Guan * and Shuaian Wang

Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Hong Kong; xuecheng-simon.tian@connect.polyu.hk (X.T.); hans.wang@polyu.edu.hk (S.W.)
* Correspondence: minguan2018@gmail.com

Abstract: Decision making under uncertainty is pivotal in real-world scenarios, such as selecting the shortest transportation route amidst variable traffic conditions or choosing the best investment portfolio during market fluctuations. In today's big data age, while the predict-then-optimize framework has become a standard method for tackling uncertain optimization challenges using machine learning tools, many prediction models overlook data intricacies such as outliers and heteroskedasticity. These oversights can degrade decision-making quality. To enhance predictive accuracy and consequent decision-making quality, we introduce a data transformation technique into the predict-then-optimize framework. Our approach transforms target values in linear regression, decision tree, and random forest models using a power function, aiming to boost their predictive prowess and, in turn, drive better decisions. Empirical validation on several datasets reveals marked improvements in decision tree and random forest models. In contrast, the benefits of linear regression are nuanced. Thus, while data transformation can bolster the predict-then-optimize framework, its efficacy is model-dependent. This research underscores the potential of tailoring transformation techniques for specific models to foster reliable and robust decision-making under uncertainty.

Keywords: uncertain decision making; predict-then-optimize; data transformation; data-driven optimization

MSC: 90-10

check for updates

Citation: Tian, X.; Guan, Y.; Wang, S. Data Transformation in the Predict-Then-Optimize Framework: Enhancing Decision Making under Uncertainty. *Mathematics* **2023**, *11*, 3782. https://doi.org/10.3390/math11173782

Academic Editors: Yu-Wang Chen, Mi Zhou and Tao Wen

Received: 10 August 2023
Revised: 1 September 2023
Accepted: 2 September 2023
Published: 3 September 2023

1. Introduction

Decision making under uncertainty is ubiquitous in real life. Real-world applications often involve making decisions with inherent uncertainty [1,2], such as the shortest transportation route when faced with uncertain and different traffic conditions [3,4] or choosing the best investment portfolio amidst market volatility [5,6]. These problems typically require selecting the best solution from multiple candidate options under uncertainty. To solve these problems [7–9], we need to consider the impact of the uncertainty on downstream decisions.

Given the complexity of current uncertain decision-making problems and the increasing volume in data, solving those problems requires advanced data analytics methods, such as machine learning tools [10]. A commonly used approach is the predict-then-optimize framework [11,12]. First, the prediction model is constructed based on the mapping between the relevant features and the uncertain parameter in the optimization model. Then, the optimization model makes the decision based on the predicted data attained from the prediction model.

For simplicity, this paper focuses on continuous uncertain parameters, in which regression models are used for prediction. The prediction model [13,14] can be represented by $y = f(x_1, x_2, \ldots, x_k) + \varepsilon$, where f represents the prediction function and $(x_1, x_2, \ldots, x_k)$ is the input vector denoting the features used to predict the response variable y, whose goal is to find the appropriate parameters such that the error ε between the predicted value $f(x_1, x_2, \ldots, x_k)$ and the true value y is as small as possible.

Various techniques, such as the deep learning model [15,16] and support vector regression model [17,18], have been used in predicting uncertain parameters in decision-making problems. Decision-making problems in the real world face complex and diverse situations; nevertheless, these traditional studies have often overlooked the impact of the intricacy of data, outliers [19], and heteroscedasticity [20] on the decision-making process. This oversight may result in sub-optimal decision when addressing practical problems.

Among these influencing factors, heteroskedasticity, as a widespread phenomenon, has a significant impact on predictive performance, which means that the variance of the explanatory variables changes as the response variable changes. This phenomenon may cause the prediction model to perform well in some regions of the data and poorly in others, thus reducing the decision quality of predicted parameters. For example, in the investment portfolio selection process, we would like to predict stock prices by using some economic indicators. However, due to market instability, stock prices may fluctuate significantly differently during different market phases, resulting in the variance of the error term varying with the state of the market [21]. Consequently, an investor might underestimate the risks associated with certain stocks in turbulent times, resulting in overexposure to volatile assets and potential financial losses. Outliers are observations that are significantly different from the majority of data points, and they may be due to measurement errors or chance circumstances. Outliers may interfere with the quality of the model's predictions, thereby reducing the accuracy of the model. For example, in the house selection process, we would like to predict the price of a house from its size and other characteristics. However, if the dataset contains some outliers, such as records of extremely high- or low-priced home transactions, these outliers may cause the model to make large errors in its predictions [22]. Such misconceptions could lead to overpayments, missed investment opportunities, or misguided selling prices based on the distorted data.

To compensate for the above shortcomings of predictive models as well as to improve the decision-making performance using predictive parameters, we utilize the data transformation technique to optimize the predictive model in the predict-then-optimize framework. Data transformation methods can help reduce heteroskedasticity by transforming the response variables in the model so that the variance of the data becomes more homogeneous. At the same time, it can effectively reduce the proportion of outliers in the data and reduce the impact of outliers on the predictive model. Consequently, using the characteristics of the data and the attributes of the decision-making problem, the appropriate data transformation method is selected to improve the accuracy of the predictive model and further enhance the quality of decision making.

Our research specifically concentrates on three common prediction models: the linear regression model, the decision tree model, and the random forest model. We convert their target values y in the three models via the power function. This transformation method aims to improve the predictive performance of machine learning models, leading to better real-world decision-making abilities under uncertainty. Via experimental validation using several real-world datasets, we find that with decision tree and random forest models, the data transformation technique can effectively improve both the predictive and decision-making quality. With linear regression models, their impact is limited. Therefore, it is feasible to use data transformation techniques to improve the performance of the predict-then-optimize framework, but it is necessary to select data transformation schemes based on different models.

In this study, we delve into the challenges of decision making under uncertainty by using the predict-then-optimize framework. Central to our work is the enhancement of both prediction and decision accuracy via data transformation. Our primary contribution is the innovative integration of data transformation techniques with diverse prediction models to bolster their efficacy in deriving decisions. The essence of our methodology is to augment decision performance by synergizing various models with data transformation. This approach offers fresh perspectives and methodologies for studies addressing analogous challenges.

In the remainder of this paper, we introduce the background and research significance of the decision-making problems and explore the application of prediction models to these problems. Subsequently, we apply data transformation techniques to three common prediction models. Finally, we conduct experiments on several real-world datasets to verify the generalizability of the proposed method.

2. Related Literature

Decision making is a crucial research area focusing on identifying the optimal solution from multiple options. Its applications are vast: in manufacturing, it helps in choosing raw materials to cut production costs [23]; in logistics, it aids in selecting optimal transportation routes to reduce costs and time [24]; and in the medical sector, it streamlines the creation of treatment programs, enhancing treatment efficacy and patient satisfaction [25].

Among decision problems, the “selection problem” has received a considerable amount of attention. In this study, we focus on this particular class of decision problems. In real-world decision-making scenarios, we often need to select the best solution among many alternatives. This choice not only impacts resource utilization but also determines the decision-making quality.

When solving uncertain selection problems, most studies usually adopt a two-stage framework known as the predict-then-optimize framework. This framework involves creating a regression or classification model for predictions and plugging these predictions into the downstream optimization model to decide on final results. When regression or classification models are constructed, uncertain parameters can be predicted based on multiple features to aid in the decision-making process. Srivastava et al. [26] draw on machine learning methods to construct mathematical models for heart disease risk prediction and to assist medical staff in medical diagnosis. Wang et al. [27] predict the risk of corporate finance risk with the help of the Light Gradient Boosting Machine method to improve the efficiency of corporate finance. Wei et al. [28] present a predictive model for pipeline safety assessment based on the eXtreme Gradient Boosting algorithm, applying grid searches to optimize the model. The construction of the model can significantly reduce the downsizing of manpower and physical resources in non-destructive examination and engineering assessment.

Although many studies have started leveraging machine learning tools for data-driven optimization, they often neglect the complexity of real-world data during the prediction stage. One such oversight is heteroskedasticity during model training. This can cause models to be overly sensitive to data performance under specific conditions, leading to unstable decision making. For instance, Lee et al. [29] find that heteroskedasticity affects users’ choices of travel modes between cities. Models that do not take heteroskedasticity into account fall short in supporting users’ travel modes choices. Similarly, Morgan [30] demonstrates that stock returns are heteroskedastic, thus influencing investment choices. Furthermore, outliers may influence the model training process and lead to over-sensitization of the model. Di Bella et al. [31], using data from a refinery’s sulfur recovery unit, and Kalisch et al. [32], via experiments on an ensemble of 100 semi-artificial time series, both find that properly addressing data outliers enhances model accuracy. Nevertheless, these studies primarily emphasize predictive analytics without thoroughly examining how predictive performance influences subsequent decision-making outcomes.

In real-world decision-making problems, neglecting issues such as data outliers and heteroskedasticity can compromise model performance, leading to sub-optimal decisions. Recognizing this, this study aims to refine the predictive model by considering the impact of the data structure on decision-making performance. To this end, we introduce the data transformation technique into the prediction model. By changing the distribution of the data via the transformation of the response variable y, we mitigate the effect of heteroskedasticity and outliers. This adaptation not only ensures the prediction model aligns better with data nuances but also bolsters decision-making efficacy.

3. Methods

In this section, we utilize the response variable transformation technique to optimize three prediction models. For simplicity, this paper mainly focuses on the continuous response variable and, therefore, studies regression prediction models. Then, the optimization model is constructed to perform target selection based on the uncertain parameters predicted by the prediction model.

3.1. Problem Setting

In the selection problem, the objective of the decision-making process is to minimize the cost function denoted as $c(z, y)$ via the choice of the decision variable $z \in \mathcal{Z} \subset \mathbb{Z}^{d_z}$, where $y \in \mathcal{Y} \subset \mathbb{Y}^{d_y}$ is the uncertain parameter. Additionally, contextual information, represented as $x \in \mathcal{X} \subset \mathbb{X}^{d_x}$, is related to the parameter y. Assuming that we obtain the new observation data $x_0 \in \mathcal{X} \subset \mathbb{X}^{d_x}$, the formulation of the optimal decision $z^*(x_0)$ for the optimization problem is presented as follows:

$$z^*(x_0) \in \underset{z \in \mathcal{Z}}{\operatorname{argmin}}\, \mathbb{E}_y[c(z, y)|x = x_0]. \tag{1}$$

A historical dataset $D_H = \{(x_i\ , y_i)\}_{i=1}^{|D_H|}$, where $|D\ _H|$ means the total number of samples in the dataset D_H, is available to address the problem (1). With this dataset at hand, the predict-then-optimize framework initially constructs a prediction model P to forecast the value of y based on the new observation information x_0, denoted as $\hat{y}_0$. Subsequently, this framework adopts the optimization model O, plugging $\hat{y}_0$ into the problem (1) by solving $\underset{z \in \mathcal{Z}}{\min} c(z, \hat{y}_0)$ to obtain the optimal decision solution. The pseudocode of the predict-then-optimize algorithm is shown in Algorithm 1.

Algorithm 1 The predict-then-optimize framework

```
1: Input: training dataset $D_H$, prediction model $P$, and optimization model $O$
2: Output: solution $z^*(x_0)$
3: Predicted values $\hat{y}_0 = P(x_0|D_H)$
4: Solution $z^*(x_0) = \underset{z \in \mathcal{Z}}{\operatorname{argmin}} c(z, \hat{y}_0)$
5: Return $z^*(x_0)$
```

3.2. Prediction Models

3.2.1. Linear Regression Model with the Response Variable Transformation

The linear regression model [33] is a common method in regression analysis that fits a linear relationship between an input feature vector x and a target value y via the least squares function. The goal is to find a straight line (or hyperplane in high-dimensional space) such that the predicted value of y' is as close to the actual value of y as possible.

In a linear regression model [34], it is assumed that the predicted target value y_i' is obtained by linearly transforming the input value x. For example, for one training dataset $D = \{x_i, y_i\}_{i=1}^{|D|}$, where x_i is the ith input feature vector, y_i is its target value, and $|D|$ represents the number of samples in D; y_i' is calculated as follows:

$$y_i' = w^T x_i + b, \tag{2}$$

where w is the coefficient vector to be trained by the model, and b is the error term to be obtained, which can be calculated via the least squares function, shown as follows:

$$\begin{aligned}(w^*, b^*) &= \underset{(w,b)}{\operatorname{argmin}} \sum\nolimits_{\{i|(x_i,y_i)\in D\}} \left(y_i' - y_i\right)^2 \\ &= \underset{(w,b)}{\operatorname{argmin}} \sum\nolimits_{\{i|(x_i,y_i)\in D\}} \left(w^T x_i + b - y_i\right)^2.\end{aligned} \tag{3}$$

After obtaining w^* and b^*, we could calculate the predicted value $y_0' = (w^*)^T x_0 + b^*$ for a new observation x_0.

However, in the actual model training process, when a linear regression model is used to fit the regression to the target value y, the relationship between the explanatory variable (feature) vector x and the response variable y may not be linear, showing a nonlinear trend. When the response variable y is transformed, the linear regression model can be made more suitable for predicting the target value y. Thus, the method of data transformation can cope with the abnormal values or outliers in the dataset, reduce the influence of abnormal values on the prediction model, and improve the robustness of the model. According to the characteristics of the data and the requirements of the model, the model prediction accuracy can be improved by adopting appropriate data transformation methods.

Considering that the response variable y may contain the value 0, we perform a data transformation of the response variable y using the form of a power function, which transforms the response variable y into its power square, i.e., it is processed in the form of y^{m_1}, where m_1 is the parameter of the power square for the linear regression model. Thus, the goal of the model construction process is to minimize the error between the transformed response variable y^{m_1} with its corresponding transformed prediction value $(y')^{m_1}$. Therefore, the calculation of w^* and b^* is transformed into (4), shown as follows:

$$\begin{aligned}(w^*, b^*)' &= \underset{(w,b)}{\operatorname{argmin}} \sum\nolimits_{\{i|(x_i,y_i)\in D\}} \left(\left(y_i'\right)^{m_1} - \left(y_i\right)^{m_1}\right)^2 \\ &= \underset{(w,b)}{\operatorname{argmin}} \sum\nolimits_{\{i|(x_i,y_i)\in D\}} \left(\left(w^T x_i + b\right)^{m_1} - \left(y_i\right)^{m_1}\right)^2.\end{aligned} \tag{4}$$

here, the hyperparameter m_1 is chosen based on the current model requirements. Our objective is to get better decision performance based on the predicted value and the characteristics of the dataset D.

3.2.2. Decision Tree Model with the Response Variable Transformation

The decision tree model [35] is a common method used in machine learning algorithms for establishing mapping relationships between input features and output response variables. The decision tree [36] constructs a tree structure by recursively dividing the dataset into subsets and selecting the optimal features and feature value at each node. For regression tasks, the model forecasts continuous values, generally by computing the average of the target variable values present in the corresponding leaf node.

In the decision tree construction process, first, we need to select the evaluation criteria of node splitting, for which the mean square error (MSE) is the most commonly used metric for regression tasks, calculated as follows:

$$MSE_{fs}^{D} = \frac{1}{|D|} \sum\nolimits_{\{i|(x_i,y_i)\in D\}} \left(g_{fs}(x_i) - y_i\right)^2, \tag{5}$$

where D is the current training dataset and $|D|$ means the number of samples in D, and $g_{fs}(x_i)$ is the predicted value of x_i from the decision tree model based on splitting feature f and the splitting feature value s, and MSE_{fs}^{D} is the MSE value under f and s.

Based on the MSE metric, the decision tree method traverses the current features and its corresponding feature values and selects the best feature and best feature value that reach the minimum MSE value for node splitting.

However, during the construction of the decision tree, we have multiple issues, i.e., heteroskedasticity and outliers, that may affect the predictive performance and stability of the model. To better solve these problems, we plan to optimize the decision tree model by using data transformation. Specifically, we use the power function to convert the predicted response variable y into the form of y^{m_2}, where m_2 is the parameter of the power function for the decision tree model. This data transformation can mitigate the effect of heteroskedasticity and make the variance more stable, thus improving the predictive performance and stability of the model. In addition, for outliers, power function data transformation may also reduce their impact on the model to some extent, making it more robust.

Accordingly, the calculation of the MSE is converted into:

$$\left(MSE_{fs}^{D}\right)' = \frac{1}{|D|}\sum\nolimits_{\{i|(x_i,y_i)\in D\}}\left(\left(g_{fs}(x_i)\right)^{m_2} - (y_i)^{m_2}\right)^2. \tag{6}$$

3.2.3. Random Forest Model with the Response Variable Transformation

The random forest algorithm [37] is an ensemble learning method, which is constituted by multiple decision trees. The prediction result is gained by integrating the predictions of each decision tree [38]. For regression problems, the final predicted value of the random forest algorithm for a new observation x_0 is the average of the predicted values of all decision trees, shown as follows:

$$y_0' = \frac{1}{N}\sum\nolimits_{j=1}^{N} T_j(x_0), \tag{7}$$

where N is the number of decision trees in the random forest model and $T_j(x_0)$ $(j \in \{1,\ldots,N\})$ is the predicted valued of the jth decision tree model for the new observation.

The random forest method adopts the bootstrap sampling method to take samples from the original training dataset, forming a new training subset for constructing one decision tree, which can increase the diversity of the model and improve the generalization ability of the random forest model.

Similar to the optimization of the decision tree model, we optimize the random forest algorithm using the response variable transformation technique, which converts the predicted response variable y into y^{m_3} to improve the model's ability to fit the heteroskedasticity of the data as well as to enhance the robustness of the model to outliers, where m_3 is the parameter of the power function for the decision tree model.Ppecifically, because the random forest algorithm is composed of multiple decision trees, we optimize each decision tree in the algorithm when we optimize the random forest with the data transformation, i.e., we use Equation (5) to optimize each decision tree in the algorithm.

3.3. Optimization Model

Specifically to examine the impact of predictions on the downstream decisions, this paper adopts the selection problem, which is ubiquitous in real life. In the selection problem, we assume that we make selections based on the predicted values obtained from the transformed predictive models described above. We assume that we select the largest u values from U predictions, which can be formulated as follows:

$$\max_{c}\sum\nolimits_{i=1}^{U} c_i y_i' \tag{8a}$$

$$\text{s. t.} \sum\nolimits_{i=1}^{U} c_i \leq u \tag{8b}$$

$$c_i \in \{0,1\},\ \forall i \in \{1,\ldots,U\}. \tag{8c}$$

where y'_i is the ith ($i \in \{1,\ldots,U\}$) predicted value obtained via the regression model, and c_i is the binary variable that indicates whether the ith term is chosen ($c_i = 1$) or not ($c_i = 0$). The goal of the optimization model is to maximize the total predicted number and meet the requirements such that the number of selected terms is no more than u.

4. Evaluation

In this section, we conduct computational experiments on several cases to assess the validity and robustness of the models presented in Section 3. Specifically, Section 4.1 describes the experimental setup in this paper. Subsequently, we evaluate and compare the decision quality before and after the transformation of the prediction model in Section 4.2.

4.1. Experiment Settings

In our experiments, we evaluate the robustness of our proposed models using two datasets to test their performance. The first dataset (denoted by D_d), derived from the "scikit-learn" library (https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html, accessed on 1 August 2023), is the diabetes dataset and contains 442 data records. The dataset D_d includes several physiological indicators of diabetic patients, with a target value of y as a quantitative measure of disease progression one year after baseline, used for the prediction of the extent of disease progression. There are ten features: age, sex, body mass index, average blood pressure, and six blood serum measurements. This dataset is utilized extensively to evaluate the performance of regression models. Therefore, the selection problem is that we select u_d high-risk patients from U_d patients, based on the dataset D_d.

The second dataset, denoted as D_s, is derived from port state control records of the port of Hong Kong from January 2015 to December 2019, a total of 3026 entries, sourced from the Asia-Pacific Computerized Information System (https://apcis.tmou.org/public/, accessed on 1 May 2022). Based on the literature [39,40], we consider 13 features that are closely related to ship conditions, which are ship age, gross tonnage, length, depth, beam, ship type, the total detention times of the ship, the total number of flag changes, the total number of casualties in last 5 years, the total number of deficiencies in the last inspection, and flag performance, recognized organization performance, and company performance in the Tokyo MoU. We construct our prediction model using dataset D_s to predict the deficiency counts of ships, where we perform ship selection based on the predicted values. Specifically, we prioritize the inspection of ships with larger predicted deficiencies, which helps to identify substandard vessels effectively. Therefore, the selection problem is that we select u_s high-risk vessels from U_s vessels, based on the dataset D_s.

For prediction model training, we divide the dataset into training and testing sets in the proportion of 4 : 1. To better evaluate our proposed models, we process the test set in a batch method, which is divided into multiple subsets to better simulate the framework structure of the current selection problem. Taking the ship dataset D_s as an example, we divide the training set D_s^T and the test set D_s^t according to the proportion and, for model evaluation, divide the test set D_s^t into multiple subsets, $Q_{D_s^t} = \left\{Q^1_{D_s^t}, Q^2_{D_s^t}, \ldots, Q^{k_{D_s^t}}_{D_s^t}\right\}$, $k_{D_s^t} = \left\lfloor \frac{|D_s^t|}{U_s} \right\rfloor$ where $Q^d_{D_s^t} = \left\{\left(x_{(d-1)U_s+1}, y_{(d-1)U_s+1}\right), \left(x_{(d-1)U_s+2}, y_{(d-1)U_s+2}\right), \ldots, (x_{dU_s}, y_{dU_s})\right\}$, $d \in \left\{1, \ldots, k_{D_s^t}\right\}$ and $(x,y) \in D_S^t$. We use the sum of the real value of target ships $P_{D_s^t}$ to evaluate the decision performance $P_{D_s^t}$, the predictive errors $MSE_{D_s^t}$ and $MAE_{D_s^t}$ to evaluate the predictive performance, and the coefficient of determination $R^2_{D_s^t}$, shown as follows, respectively:

$$P_{D_s^t} = \sum_{q \in Q_{D_s^t}} \sum_{\{i | (x_i, y_i) \in O(u_s, q)\}} y_i\ , \tag{9a}$$

$$MSE_{D_s^t} = \sum_{q \in Q_{D_s^t}} \sum_{\{i|(x_i,y_i) \in O(u_s,q)\}} \left(y_i' - y_i\right)^2 , \tag{9b}$$

$$MAE_{D_s^t} = \sum_{q \in Q_{D_s^t}} \sum_{\{i|(x_i,y_i) \in O(u_s,q)\}} \frac{\left|y_i' - y_i\right|}{u_s} , \tag{9c}$$

$$R^2_{D_s^t} = 1 - \frac{\sum_{i=1}^{U_s} \left(y_i - y_i'\right)^2}{\sum_{i=1}^{U_s} \left(y_i - \overline{y}\right)^2}, \overline{y} = \frac{1}{U_s} \sum_{i=1}^{U_s} y_i , \tag{9d}$$

where $O(u_s, q)$ means the set of selected ships whose predicted values are among the highest u_s ships in the ship set q. To accurately compare the prediction accuracy of the regression models before and after using the data transformation method, we select the initial MSE metric. $R^2_{D_s^t}$ is used to measure the extent to which the regression model explains the variability of the dependent variable. The closer the $R^2_{D_s^t}$ value is to 1, the better the model explains the variability of the dependent variable, representing a better fit. If the $R^2_{D_s^t}$ is close to 0, it represents that the model does not explain the variability of the dependent variable well, and the fit is poor.

To streamline our experiments, we assume that the diabetic patient selection problem is to choose two patients from a pool of four, namely $U_d = 4, u_d = 2$, and choose four patients from a pool of ten, namely $U_d = 10, u_d = 4$, who would have a more severe health condition one year later. The ship selection problem is to choose three high-risk ships from a pool of ten, namely $U_s = 10, u_s = 3$, and select three high-risk ships from a pool of twenty, namely $U_s = 20, u_s = 3$. Validation and testing in different application scenarios and instance scales under different datasets can better prove the effectiveness of our proposed technique and enhance the trustworthiness of our method in practical applications.

4.2. Evaluation of Models

In our research, we try to optimize the performance of regression models, especially for the possible issues of heteroskedasticity and outliers, which may reduce the accuracy and stability of the prediction model. To optimize prediction models and make better decisions using predictions, we use the data transformation method, in which the target variable is converted from the original y to the form of y^m. For the three prediction models presented in Section 3, we adopt the cross-validation method to find the appropriate parameter m.

First, the value of the hyperparameter m affects the complexity of constructing the model; considering the model complexity, prior experience, and practical considerations, we select the values of $m_1, m_2,$ and m_3 from set $\{1/4,\ 1/3,\ 1/2,\ 1,\ 2,\ 3\}$. In linear regression, the coefficients w and intercept b are iteratively adjusted using the gradient descent technique, and the gradient of w in the linear model can be computed as outlined below:

$$\frac{\partial loss\left(\left(y'\right)^{m_1},\ \ y^{m_1}\right)}{\partial w} = \frac{1}{|D|} \sum_{\{i|(x_i,y_i) \in D\}} \left((wx_i + b)^{m_1} - (y_i)^{m_1}\right) \times m_1 \times x_i (wx_i + b)^{m_1 - 1}, \tag{10}$$

where D means the current training data used for linear regression model and $loss\left(\left(y'\right)^{m_1}, y^{m_1}\right)$ means the MSE between the predicted values $\left(y_i'\right)^{m_1}$ and the target values $(y_i)^{m_1}$. Because the predicted values calculated via $wx_i + b$ may contain 0, the value of m must be greater than or equal to 1. Therefore, the range of values of m_1 in linear regression model is $\{1,\ 2,\ 3\}$.

Then, the cross-validation method is utilized to evaluate the model for each m value. Specifically, we divide the training dataset D^T into several parts and then sequentially use each part as the validation set and the rest as the training set and compute the decision performance of the model on the validation set as the evaluation index, which is calculated via Equation (9a). We search for the value of m that makes decision performance optimal via the cross-validation approach and evaluate in detail the performance of these models in terms of MSE and decision performance (P_{D^t}) metrics in the test dataset D^t.

We construct three different regression models in this experiment, linear regression model, decision tree model, and random forest model, optimizing those regression models via the data transformation method. We first train and test the three models on the original data to obtain their benchmark performance. Then, we transform the response variables y of the three models with power functions, respectively, and train and test them under the optimization method.

The overall results are represented in Tables 1–4, where Tables 1 and 2 show the results for D_d and Tables 3 and 4 show the results for D_s. Via the comparative analysis of results before and after transforming the response variable, we desire to gain a deeper understanding of the strengths and weaknesses of different models and provide a reliable basis for selecting the optimal regression model.

Table 1. The prediction quality and decision performance of different prediction methods for D_d when $U_d = 4,\ u_d = 2$.

Model	Optimal m	MSE		P_{D_d}		MAE		R^2	
		Before	After	Before	After	Before	After	Before	After
Linear regression	$m_1 = 1$	832,465.15	832,465.15	3982.00	3982.00	1991.00	1991.00	−4.437	−4.437
Decision tree	$m_2 = 1/2$	164,131.00	161,589.00	3890.00	4021.00	774.50	782.50	−0.318	−0.242
Random forest	$m_3 = 2$	88,552.13	86,156.87	4039.50	4044.00	566.43	561.75	0.246	0.264

Table 2. The prediction quality and decision performance of different prediction methods for D_d when $U_d = 10,\ u_d = 4$.

Model	Optimal m	MSE		P_{D_d}		MAE		R^2	
		Before	After	Before	After	Before	After	Before	After
Linear regression	$m_1 = 1$	331,327.38	331,327.38	1532.25	1532.25	383.06	383.06	−4.123	−4.123
Decision tree	$m_2 = 1/2$	74,373.25	70,775.75	1449.75	1551.50	163.69	162.56	−0.296	−0.225
Random forest	$m_3 = 2$	36,072.06	34,296.20	1530.00	1564.00	108.77	107.63	0.229	0.245

Table 3. The prediction quality and decision performance of different prediction methods for D_s when $U_s = 10,\ u_s = 3$.

Model	Optimal m	MSE		P_{D_d}		MAE		R^2	
		Before	After	Before	After	Before	After	Before	After
Linear regression	$m_1 = 1$	3070.09	3070.09	262.67	262.67	87.54	87.54	−0.757	−0.757
Decision tree	$m_2 = 1/3$	2885.41	2857.03	328.33	339.00	99.07	101.19	−0.326	−0.293
Random forest	$m_3 = 1/2$	1442.53	1423.77	396.00	397.33	67.26	66.51	0.298	0.303

Table 4. The prediction quality and decision performance of different prediction methods for D_s when $U_s = 20,\ u_s = 3$.

Model	Optimal m	MSE		P_{D_d}		MAE		R^2	
		Before	After	Before	After	Before	After	Before	After
Linear regression	$m_1 = 1$	2331.20	2331.20	165.00	165.00	55.00	55.00	−0.744	−0.744
Decision tree	$m_2 = 1/3$	2428.06	2243.19	185.67	216.67	70.17	68.24	−0.321	−0.288
Random forest	$m_3 = 1/2$	1145.56	1143.81	255.33	255.67	42.86	43.21	0.294	0.298

As shown in Tables 1–4, we found that the regression models after data transformation have improved both their decision performance (P_D), coefficient of determination (R^2), and predictive accuracy (MSE and MAE) in most models, with the exception of linear regression.

Zooming in on Table 1, for the task of selecting high-risk patients (two out of four), the decision tree method exhibits a decrease in MSE by roughly 1.5% and an increment in MAE by about 1.0%. Decision performance improves by 3.4%. Notably, while the initial R^2 is less than 0 (indicating potential issues with prediction), after data transformation, there is an increase in R^2 of approximately 23.8%, signifying improvement. The random forest model shows a decline in MSE of around 2.7%, a reduction in MAE of roughly 0.8%, a slight improvement in decision performance of 0.1%, and a boost in R^2 of about 7.3%.

As shown in Table 2, for the task of selecting high-risk patients (four out of ten), the decision tree reduces the MSE by about 4.8% and reduces a decrease in MAE by roughly 0.6%, whereas the decision performance improves by 7.0%, and R^2 is increased by about 23.9%. The random forest exhibits a decrease in MSE of about 4.9% and reduces MAE by roughly 0.9%. Decision performance improves by 2.2% and R^2 increases by approximately 6.9%.

Meanwhile, for the high-risk ship selection problem of three out of ten, shown in Table 3, the decision tree's prediction quality of MSE increases by roughly 0.9% and the increment in MAE by approximately 2.0%, while the decision performance improves by 3.2% and there is an increase in R^2 of approximately 10.0%. For the random forest model, its prediction accuracy of MSE is improved by about 1.3%, and the MAE is reduced by about 1.5%. Decision performance goes up by about 0.3%, and R^2 is increased by about 1.7%.

The results in Table 4 are about selecting three ships from a pool of twenty. The decision tree shows a decline in MSE of roughly 7.6% and reduces the MAE by approximately 2.8%, whereas the decision performance improves by 16.7% and R^2 increases by about 10.2%. The random forest shows a decline in MSE of roughly 0.2% and a boost in MAE of about 2.3%. Decision performance improves by approximately 0.1% and R^2 is enhanced by about 1.3%.

4.3. Discussion and Analysis

Combining the experimental results from the two selection problems, the linear regression model does not improve its performance after the response variable transformation. This likely stems from the model's inability to adeptly fit the transformed data given the dataset's characteristics. However, the transformation technique is more effective for the decision tree and the random forest, probably because these two models can better capture and construct the nonlinear relationships in the two current datasets.

The linear regression method is highly sensitive to the data distribution and requires checking for linearity in regression tasks. Consequently, data transformation requirements are more stringent, often necessitating multiple tests to find a suitable transformation method. On the other hand, the decision tree method is less sensitive to data distribution, allowing for the use of data transformation to mitigate the effects of outliers and heteroskedasticity. Random forest, comprising multiple decision trees, shares similarities with the decision tree model. Selecting the appropriate data transformation method involves considering factors such as data distribution and expertise knowledge. The most effective approach is to experiment with different methods and evaluate their performance to determine the optimal transformation method for a specific model and application.

In summary, this study shows the impact of response variable transformation on different algorithms and datasets, whose results are slightly different. In general, the data transformation method can effectively improve the model's prediction and decision performance.

5. Conclusions

In this study, we focus on optimizing prediction models under the predict-then-optimize framework, so as to improving the decision-making performance for the downstream optimization problems. To this end, we adopt the response variable transformation technique. Our research is based on the three common prediction models, i.e., linear regression model, decision tree model, and random forest model. After performing data transformation on the three regression models, we obtain some interesting results. For the decision tree and random forest models, after optimization, both their decision performance and prediction quality are improved. That suggests that data transformation is effective in these models and can improve decision-making performance. However, for the linear model, there is no significant effect. This may be due to the characteristics of the model itself, which may be more suitable for dealing with linear relationships, and the nonlinear features introduced by data transformation are not helpful for model performance. In conclusion, using data transformation to improve prediction models is feasible under the traditional predict-then-optimize framework, but it is necessary to choose appropriate data transformation methods according to different models and applications. These findings are instructive for better solving decision-making problems and provide references for future research.

Author Contributions: Conceptualization, X.T. and S.W.; methodology, S.W.; software, Y.G.; validation, Y.G.; formal analysis, Y.G.; investigation, Y.G.; resources, S.W.; data curation, S.W.; writing–original draft preparation, X.T. and Y.G.; writing–review and editing, X.T. and S.W.; supervision, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Available if requested.

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

References

1. Byrnes, J.P. The development of decision-making. *J. Adolesc. Health* **2002**, *31*, 208–215. [CrossRef] [PubMed]
2. Simon, H.A.; Dantzig, G.B.; Hogarth, R.; Plott, C.R.; Raiffa, H.; Schelling, T.C.; Shepsle, K.A.; Thaler, R.; Tversky, A.; Winter, S. Decision making and problem solving. *Interfaces* **1987**, *17*, 11–31. [CrossRef]
3. Wang, Y.; Peng, S.; Zhou, X.; Mahmoudi, M.; Zhen, L. Green logistics location-routing problem with eco-packages. *Transp. Res. Part E Logist. Transp. Rev.* **2020**, *143*, 102118. [CrossRef]
4. Pečený, L.; Meško, P.; Kampf, R.; Gašparík, J. Optimisation in transport and logistic processes. *Transp. Res. Procedia* **2020**, *44*, 15–22. [CrossRef]
5. Shanmuganathan, M. Behavioural finance in an era of artificial intelligence: Longitudinal case study of robo-advisors in investment decisions. *J. Behav. Exp. Financ.* **2020**, *27*, 100297. [CrossRef]
6. Vo, N.; He, X.; Liu, S.; Xu, G. Deep learning for decision making and the optimization of socially responsible investments and portfolio. *Decis. Support Syst.* **2019**, *124*, 113097. [CrossRef]
7. Liu, Z.; Wang, Y. Handling constrained multiobjective optimization problems with constraints in both the decision and objective spaces. *IEEE Trans. Evol. Comput.* **2019**, *33*, 870–884. [CrossRef]
8. Shabani, A.; Asgarian, B.; Salido, M.; Gharebaghi, S.A. Search and rescue optimization algorithm: A new optimization method for solving constrained engineering optimization problems. *Expert Syst. Appl.* **2020**, *161*, 113698. [CrossRef]
9. Bérubé, J.; Gendreau, M.; Potvin, J. An exact ϵ-constraint method for bi-objective combinatorial optimization problems: Application to the Traveling Salesman Problem with Profits. *Eur. J. Oper. Res.* **2009**, *194*, 39–50. [CrossRef]
10. Xu, Y. Data-Driven Dynamic Decision Making: Algorithms, Structures, and Complexity Analysis. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2023.
11. Wang, S.; Yan, R.; Qu, X. Development of a non-parametric classifier: Effective identification, algorithm, and applications in port state control for maritime transportation. *Transp. Res. Part B Methodol.* **2019**, *128*, 129–157. [CrossRef]
12. Yang, Z.; Yang, Z.; Yin, J. Realising advanced risk-based port state control inspection using data-driven Bayesian networks. *Transp. Res. Part A Policy Pract.* **2018**, *110*, 38–56. [CrossRef]
13. Grömping, U. Variable importance in regression models. *Wiley Interdiscip. Rev. Comput. Stat.* **2015**, *7*, 137–152. [CrossRef]
14. Fitzmaurice, G. Regression. *Diagn. Histopathol.* **2016**, *22*, 271–278. [CrossRef]
15. Akita, R.; Yoshihara, A.; Matsubara, T.; Uehara, K. Deep learning for stock prediction using numerical and textual information. In Proceedings of the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 26–29 June 2016; pp. 1–6. [CrossRef]
16. Zhu, W.; Xie, L.; Han, J.; Guo, X. The application of deep learning in cancer prognosis prediction. *Cancers* **2020**, *12*, 603. [CrossRef] [PubMed]
17. Sun, Y.; Ding, S.; Zhang, Z.; Jia, W. An improved grid search algorithm to optimize SVR for prediction. *Soft Comput.* **2021**, *25*, 5633–5644. [CrossRef]

18. Panahi, M.; Sadhasivam, N.; Pourghasemi, H.R.; Rezaie, F.; Lee, S. Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). *J. Hydrol.* **2020**, *588*, 125033. [CrossRef]
19. Peña, D. Detecting outliers and influential and sensitive observations in linear regression. In *Springer Handbook of Engineering Statistics*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 605–619. [CrossRef]
20. Tan, F.; Jiang, X.; Guo, X.; Zhu, L. Testing heteroscedasticity for regression models based on projections. *Stat. Sin.* **2021**, *31*, 625–646. [CrossRef]
21. Motegi, K.; Iitsuka, Y. Inter-regional dependence of J-REIT stock prices: A heteroscedasticity-robust time series approach. *N.Am. J. Econ. Financ.* **2023**, *64*, 101840. [CrossRef]
22. Zaki, J.; Nayyar, A.; Dalal, S.; Ali, Z.H. House price prediction using hedonic pricing model and machine learning techniques. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7342. [CrossRef]
23. Meng, L.; McWilliams, B.; Jarosinski, W.; Park, H.; Jung, Y.; Lee, J.; Zhang, J. Machine learning in additive manufacturing: A review. *JOM* **2020**, *72*, 2363–2377. [CrossRef]
24. Basso, R.; Kulcsár, B.; Sanchez-Diaz, I. Electric vehicle routing problem with machine learning for energy prediction. *Transp. Res. Part B Methodol.* **2021**, *145*, 24–55. [CrossRef]
25. Poldrack, R.A.; Huckins, G.; Varoquaux, G. Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry* **2020**, *5*, 534–540. [CrossRef] [PubMed]
26. Srivastava, A.; Kumar, S.A. Heart Disease Prediction using Machine Learning. In Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 28–29 April 2022; pp. 2633–2635. [CrossRef]
27. Wang, D.; Li, L.; Zhao, D. Corporate finance risk prediction based on LightGBM. *Inf. Sci.* **2022**, *602*, 259–268. [CrossRef]
28. Liu, W.; Chen, Z.; Hu, Y. XGBoost algorithm-based prediction of safety assessment for pipelines. *Int. J. Press. Vessel. Pip.* **2022**, *197*, 104655. [CrossRef]
29. Lee, J.H.; Chon, K.S.; Park, C. Accommodating heterogeneity and heteroscedasticity in intercity travel mode choice model: Formulation and application to HoNam, South Korea, high-speed rail demand analysis. *Transp. Res. Rec.* **2004**, *1898*, 69–78. [CrossRef]
30. Morgan, I.G. Stock prices and heteroscedasticity. *J. Bus.* **1976**, *49*, 496–508. [CrossRef]
31. Di Bella, A.; Fortuna, L.; Graziani, S.; Napoli, G.; Xibilia, M.G. A comparative analysis of the influence of methods for outliers detection on the performance of data driven models. In Proceedings of the 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007, Warsaw, Poland, 1–3 May 2007; pp. 1–5. [CrossRef]
32. Kalisch, M.; Michalak, M.; Sikora, M.; Wróbel, Ł.; Przystałka, P. Influence of outliers introduction on predictive models quality. In Proceedings of the Advanced Technologies for Data Mining and Knowledge Discovery: 12th International Conference, BDAS 2016, Ustroń, Poland, 31 May–3 June 2016; pp. 79–93. [CrossRef]
33. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
34. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. Linear regression. In *An Introduction to Statistical Learning: With Applications in Python*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 69–134. [CrossRef]
35. Myles, A.J.; Feudale, R.N.; Liu, Y.; Woody, N.A.; Brown, S.D. An introduction to decision tree modeling. *J. Chemom.* **2004**, *18*, 275–285. [CrossRef]
36. Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283. [CrossRef]
37. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Yan, R.; Wang, S.; Fagerholt, K. A semi-"smart predict then optimize" (semi-SPO) method for efficient ship inspection. *Transp. Transp. Res. Part B Methodol.* **2020**, *142*, 100–125. [CrossRef]
40. Yan, R.; Wang, S.; Peng, C. An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities. *J. Comput. Sci.* **2021**, *48*, 101257. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.