Abstract: The coronavirus disease (COVID-19) pandemic has flooded public health organizations around the world, highlighting the significance and responsibility of medical crowdfunding in filling a series of gaps and shortcomings in the publicly funded health system and providing a new fundraising solution for people that addresses health-related needs. However, the fact remains that medical fundraising from crowdfunding sources is relatively low and only a few studies have been conducted regarding this issue. Therefore, the performance predictions and multi-model comparisons of medical crowdfunding have important guiding significance to improve the fundraising rate and promote the sustainable development of medical crowdfunding. Based on the data of 11,771 medical crowdfunding campaigns from a leading donation-based platform called Weibo Philanthropy, machine-learning algorithms were applied. The results demonstrate the potential of ensemble-based machine-learning algorithms in the prediction of medical crowdfunding project fundraising amounts and leave some insights that can be taken into consideration by new researchers and help to produce new management practices.

Keywords: crowdfunding; medical crowdfunding; machine learning; fundraising prediction; weibo philanthropy

1. Introduction

Pandemics are increasingly becoming a constant threat to human beings [1,2]. As the latest example, the coronavirus (COVID-19) pandemic, which has affected public healthcare systems and caused severe economic and social distress worldwide [3–5], reminded people of the significance of medical crowdfunding. As a novel and noteworthy fundraising channel, medical crowdfunding helps with relieve the high medical cost burdens [6–8]. Among one type of donation-based crowdfunding supported by an online platform, medical crowdfunding has arisen as a new form of charitable fundraising solution for individuals dealing with financial pressures and reducing the possibility of personal or family bankruptcy from the costs of medical treatment [9].

More importantly, medical crowdfunding plays an increasingly crucial complementary role when there are certain gaps and deficiencies (e.g., a lack of public insurance, difficulty in obtaining specialist care and long waiting time for treatment) in the publicly funded health system under a pandemic situation. Medical crowdfunding therefore is a good response to these inadequacies and represent a way to satisfy health-related needs [10]. The total amount of global crowdfunding platform financing reached US $72.3 billion in 2017, about half of this amount was collected from medical crowdfunding [11]. Raisers could approach the public and raise the necessary funds for their proposed aiding projects through medical crowdfunding projects [9]. Based on the continuous improvement of
“digital philanthropy” and sharing economy, medical crowdfunding has become widely accepted and now represents a channel with great potential.

During the Covid-19 pandemic, a medical crowdfunding platform called Weibo Philanthropy (Weibo GongYi) took on a great deal of social responsibility and performed very well. The platform itself first invested 100 million Yuan a special rescue fund and then opened donation channel used by early 2,800,000 people that, together, donated about approximately 8 million US dollars to support the fight against the epidemic through this platform. For example, the first project to offer support in the pandemic was achieved its target of 22,000 US Dollars in 50 min [12].

Even if relatively low socio-economic classes and under-educated groups are in urgent need for funds and ways to promote it [13], developing a high-performance fundraising project without guidance is still extremely challenging [11,14,15]. Therefore, it is applicable to predict the results of project financing through understanding the influencing factors and attributes of medical crowdfunding. With a better understanding of this scheme, it would be possible to efficiently give particular guidance and optimization to craft fundraising campaigns. This is one of the reasons why this study was conducted.

The aim of this paper is to take “Weibo Philanthropy” as the key example, and apply machine-learning algorithm to study the data of Project (including Target Amount; Funding duration; Number of updates; Word count of description; Fundraising success rate) and data of Participant (Type of raisers; Number of donors; Number of retweets). The purpose of this paper is to predict the impact of these factors on the performance of medical crowdfunding projects and the capacity of projects’ fundraising, to provide more conducive suggestions to fundraisers. This paper attempts to make the following contributions.

Firstly, the current research on donation-based crowdfunding mainly focuses on the concept, characteristics, impact factors and operation mechanism [16,17]. These attributes related to the performance of medical financing need to be further studied [11]. Other features are centered on the motivation of donors [18–21], or individual biases [22,23]. Moreover, there are few pieces of research on the topic segmentation into medical crowdfunding, let alone adopting machine learning to study this topic. For instance, if the topic of “crowdfunding” is searched on Google Scholar, it will return 6430 results. However, just 39 of those results would be for “medical crowdfunding”, and only 2 results would involve the use of machine learning for medical crowdfunding, representing a gap on the field that must be filled.

Secondly, the innovative database is also one of the leads of this study. Unlike previous research data, which is generally obtained from GoFundMe, GiveForward or Kickstarter, the data used here was crawled from the popular platform “Weibo Philanthropy” (gongyi.weibo.com), between February 2012 and March 2020 (eight years) on this platform, which has not been studied before. It is an online platform that was established in 2012 for needy individuals and groups to solicit public welfare fundraising campaigns. It also carries out grassroots activities nationwide. In the past three years, this site has raised over 400 million donations from more than 20 million users [24].

Thirdly, in reality, the success of donation-based crowdfunding projects, especially when it comes to medical assistance projects, does not entirely driven by the behavior of donors. Factors such as the raiser’s understanding of the project, the donor’s attitude towards the project and the interaction between participants, may also affect the performance of crowdfunding campaigns.

Lastly, the fundraising of medical crowdfunding is an uncertain process, and its performance changes dynamically [11,25]. Based on this, this paper dynamically predicts the financing outcomes to guide raisers how to launch and follow up projects more efficiently. With a methodology based on machine learning, it also puts forward a suggestion for a reasonable fundraising target amount. It is noteworthy that this is an exploring research, since systematic attempts to discover the potential and advantages of using machine learning for medical crowdfunding financing predictions through research are still lacking.
2. Literature Review

2.1. Donation-Based Crowdfunding

The concept of crowdfunding comes from the broader concept of crowdsourcing [26]. Such strategy focuses on the idea of using people, in other words, the crowd to obtain ideas, feedbacks and solutions to develop corporate activities [27], and it is now considered a new source of value creation [23, 28–30]. A myriad of social causes can contribute to the emersion of a donation-based crowdfunding, which generally respond to the lack of consummation of public goods. Moreover, it offers financial or social aids for individuals and communities that could hardly seek help elsewhere [31].

The prosocial motivation, defined as the desire to help others, is perceived to be most considered in donation-based crowdfunding. Individuals behave prosocially for reasons such as feeling good about helping others, making impact or developing a favorable image—all the motives usually associated with the funding performance [32, 33]. Previous researchers have suggested the people with higher engagement in prosocial behaviors tend to make financing decisions for public benefits [18, 34]. For this reason, donation-based crowdfunding is very likely to attract a wide variety of potential contributors. Notably, projects on donation-based platforms have concrete and well-defined terms and guidelines, especially in comparison with conventional and loosely constructed fundraising campaigns.

2.2. Medical Crowdfunding

As one of donation-based crowdfunding, medical crowdfunding arises as a novel solution and configurations for individuals dealing with short-term financial pressures from the costs of medical treatment. Through medical crowdfunding platform raisers could approach the crowd and raise necessary funds for their proposed aiding projects online [9]. These platforms/websites dis-intermediate the fundraising process and exert strong network effects among contributors [35]. At the same time, medical care could incorporate with be positively linked with poverty reduction and children’s education and other projects [36–39]. Medical crowdfunding is increasingly becoming an opportunity area of crowdfunding with great potential and might become an effective way to alleviate financial pressures of patients [10, 40–42]. Importantly, it differs from traditional approach in several aspects and such differences must be acknowledge.

Firstly, increased funding requirements for health initiatives, such as public health vaccine development, for instance, make medical crowdfunding as a viable funding channel [38, 40]. Individuals who are sick, especially those who cannot afford the cost for medical costs, turn to crowdfunding to raise funds [10, 15, 42, 43]. Medical crowdfunding campaigns are the channels to collect funds for various reasons: hospital costs, disease treatment, home care needs, post diagnostic agreements, general support for drugs, postoperative care. In this scenario, a number of dedicated medical crowdfunding platforms have emerged, including YouCaring, Giveforward and GofundMe in US, Weibo Philanthropy and Tecent Donation in China.

Secondly, one significant feature of medical crowdfunding is emphasizing beneficiaries’ abilities in soliciting funds in terms of the scale. Environmental trend factors, including geographical constraints, deficiency of medical insurance and health care regulations, drive population in need to medical crowdfunding platforms to get funded [10, 15, 38, 44, 45]. The scale of capital acquisition has been expanded by aggregating potential donors in nominated locations, eliminating offline spatial and temporal constraints [22, 46].

Thirdly, information sharing features are one of the greatest features of Medical crowdfunding and can be linked to a boost in quantity and speed of the fundraising. The platform not only enables funding projects to simultaneously broadcast to immediate relations such as friends and family, but also considerably propagates information sharing via social media to broader social relations, thus helping to spread the word [36]. That said, the quantity of potential donors of medical crowdfunding could easily outperforms traditional face-to-face solicitations.
Finally, the threshold of donation is extremely low, since contributor could voluntarily donate a very small amount of money. Since the medical crowdfunding network has formulated and attracted an ocean of existing and potential donors, the financial resources are accumulated and aggregated exponentially. Therefore, with a low threshold of donation, individuals are more likely to give a hand. Even those with very weak ties to the beneficiary may donate [47].

2.3. Research on Crowdfunding Prediction with Machine Learning

Previous studies on crowdfunding prediction usually apply traditional statistical methods, such as linear regression and logistic regression [48,49], assuming that the input variables are also independent variables. Consequently, the regression method may produce a larger prediction error when the input variables are correlated [50]. Although machine learning is an effective method to analyze hidden association in big data set and identify complex data [51], scholars begin to apply such algorithms to practical issues. For example, using machine-learning algorithm to predict the success of projects [50,52,53], such as Support Vector Machine(SVM), Decision Trees (DT) and K-Nearest Neighbor(KNN) algorithm [54–57], as well as using XgBoost, Gradient Boosting, Random Forest and GLM to construct prediction models [58]. Kamath et al. (2016) builds several supervised learning models to predict success rate of crowdfunding campaigns (including Neural Network, Random Forest, Naïve Bayes and Decision Tree) and found that Neural Network performed better [59]. Furthermore, a multi-modal deep learning framework also has been developed to predict the outcome of crowdfunding projects [60]. The others are based on a framework of text, and using a Random Forest algorithm to predict the financing performance of crowdfunding [61].

Notwithstanding the aforementioned novel methods, their studies are focused on the category of general donation-based crowdfunding. Based on that, the subdivision that is the focus of this paper is medical crowdfunding campaigns belonging to donation-based crowdfunding category. In addition, the prediction model here proposed is based on eXtreme Gradient Boosting algorithm and, differently than its applicability in other studies; it is used to compare it with other four algorithm (including Classification and Regression Tree, K-Nearest Neighbor, Linear Regression, Artificial Neural Network). To the best of our knowledge, this study is the first to use machine-learning algorithm to establish an effective prediction model for Weibo Philanthropy medical crowdfunding campaigns, and compared the five algorithms to get the higher accuracy outcome.

3. Methodology

3.1. Data Collection

The data for this research was crawled from a leading donation-based platform called “Weibo Philanthropy” (gongyi.weibo.com), established in 2012 with the aim to help individuals and groups in need to solicit public welfare fundraising campaigns. It is an affiliated crowdfunding platform of Sina Weibo currently the largest micro-blogging platform in China, with over 516 million active users. The platform’s impressive fundraising performance, and considerable number of active users were the main reasons why it was chosen as a subject for this research, not to mention its wide range of dissemination. Nearly 28 million people donated about $8.3427 million in medical crowdfunding projects in Weibo Philanthropy to support the fight against the COVID-19 epidemic. The first project of them reached its target of $22,000 in 50 min (Sina 2020). It raised more than 100 million Yuan in 57 h to help the victims of the Lushan earthquake in Sichuan Province (Sina 2013). Thus, it can be said that this platform provides reliable data with multiple attributes for training and testing models, which make the prediction results more realistic and more comparable. The analysis object of this work is data between February 2012 and March 2020 (eight years) on this platform, which reflect the fact on online medical crowdfunding field comprehensively.
3.2. Data Preprocessing Analysis

The data set contains 11,771 instances of medical crowdfunding projects. As shown in Figure 1, first step, 25,413 sets of donation-based crowdfunding data from February 2012 to March 2020 on the website were crawled by using Python, followed by the second step to remove 4558 projects that were still in the process of completion. Then, 1747 sets of duplicate data or messy data as well as other data sets except medical crowdfunding in donation-based crowdfunding (including Animal Protection, Education, Environment Protection, Poverty Alleviation) were taken away. Finally, the data sets with missing “Target Amount” in medical crowdfunding campaigns were cleared away. Upon completing the above five steps, we get 11,771 sets of data. As shown in Table 1 each sample consists of 8 factors about both project and participant, including type of raisers, number of donors, number of retweets, target amount, funding duration, number of updates, word count of description, and fundraising. The factors were chosen according to the literature and properly summarized in Table 2. We take the fundraising which is the total amount of raised funds as the target feature because the purpose of our work is to analyze the influencing factors of the fundraising and predict the fundraising ability of users. After selection, there are 11,771 samples left in the data set that contain 2 missing values of 2 samples. We filled both missing values of duration via the update report of these two samples, respectively. The visualized relationships between each attribute and the final amount of crowdfunding (fundraising) are shown in Figure 2 which explicates there are outliers in the data set intuitively. We calculated 24 outliers by comparing the Euclidean distance between the sample and the mean value of each attribute and setting the hyperparameters based on the principle of keeping the greatest number of objects. Figure 3, containing 11,747 samples in each subgraph, illustrates the relationship between each attribute and the final amount of crowdfunding (fundraising).

Table 1. Feature and description.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TypeRaiser</td>
<td>Type of raisers (personal = 0, organization = 1)</td>
</tr>
<tr>
<td>Target</td>
<td>The dollar amount that is sought through the campaign (CNY)</td>
</tr>
<tr>
<td>NumRetweet</td>
<td>Number of people who shared the campaign link</td>
</tr>
<tr>
<td>Duration</td>
<td>Time period for which the campaign has been active (days)</td>
</tr>
<tr>
<td>NumDonator</td>
<td>Number of people who donated to a campaign</td>
</tr>
<tr>
<td>NumUpdate</td>
<td>Number of updates between the launched time and the ended time</td>
</tr>
<tr>
<td>NumWord</td>
<td>The number of words contained in project description.</td>
</tr>
<tr>
<td>Fundraising</td>
<td>The total amount of raised fund</td>
</tr>
</tbody>
</table>

Table 2. Summary of factors and related references.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Related References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>Sudhir et al. (2016), Ryzhov et al. (2016), Gleasure and Feller (2016), Qian and Lin (2017), Sulaeman and Lin (2018), Chen et al. (2020)</td>
</tr>
<tr>
<td>Number of Retweets</td>
<td>Kromidha and Robson (2016), De Vries et al. (2012), Liu et al. (2017)</td>
</tr>
<tr>
<td>Number of Words</td>
<td>Qian and Lin (2017), Sulaeman and Lin (2018), Zhou et al. (2018)</td>
</tr>
<tr>
<td>Number of Donors</td>
<td>Kromidha and Robson (2016), Sudhir et al. (2016), Giudici et al. (2018), Gleasure and Morgan (2018)</td>
</tr>
</tbody>
</table>
25,413 donation-based crowdfunding campaigns on Weibo Philanthropy
(From February 2012 to March 2020)

Excluded 4558 campaigns in the phase of fundraising or implementation

20,855 accomplished donation-based crowdfunding campaigns

Excluded 1747 campaigns with invalid data

19,108 donation-based crowdfunding campaigns with valid data

Excluded 7324 campaigns in other categories, such as Animal Protection, Education, Environment Protection, Poverty Alleviation and others.

11,784 accomplished crowdfunding campaigns in medical assistance category

Excluded 13 campaigns missing data of "Target amount"

11,771 accomplished medical crowdfunding campaigns with complete data are selected to form the final sample

Figure 1. Sample selection process.
The distribution of fundraising is shown in Figure 4. Obviously, it approximates a long-tailed distribution. The min, median, mean, max, and standard deviation of crowd-funding are 136,900, 605,883.25, 7,259,000, and 1,111,957.73, respectively. Table 3 shows the descriptive statistics for each factor of the data set. As illustrated in Table 3, the difference among the mean of each factor is large. However, the large differences affect the output significantly of machine-learning models. To eliminate the dimensional influence between
features, data standardization is required. Standardization is a sort of technology that scales the data to a small specific interval removing the unit limit of the data and converting data into a dimensionless value. In this way, the contribution of each feature to the result is balanced, and the iterative speed of gradient descent can also be improved. We perform Min-Max normalization converts each feature to the interval [0,1].

Table 3. Descriptive statistics.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>100</td>
<td>50,000</td>
<td>52,830.43</td>
<td>1,000,000</td>
<td>57,881.63</td>
</tr>
<tr>
<td>NumRetweet</td>
<td>0</td>
<td>34</td>
<td>123.70</td>
<td>999</td>
<td>229.29</td>
</tr>
<tr>
<td>Duration</td>
<td>0</td>
<td>60</td>
<td>56.04</td>
<td>180</td>
<td>14.76</td>
</tr>
<tr>
<td>NumDonator</td>
<td>0</td>
<td>82</td>
<td>227.82</td>
<td>14,315</td>
<td>637.63</td>
</tr>
<tr>
<td>NumUpdate</td>
<td>0</td>
<td>4</td>
<td>5.47</td>
<td>42</td>
<td>4.67</td>
</tr>
<tr>
<td>NumWord</td>
<td>66</td>
<td>1103</td>
<td>1096.78</td>
<td>2713</td>
<td>409.41</td>
</tr>
<tr>
<td>Fundraising</td>
<td>0</td>
<td>136,900</td>
<td>605,883.25</td>
<td>7,259,000</td>
<td>1,111,957.73</td>
</tr>
</tbody>
</table>

Figure 4. Distribution of Fundraising.

3.3. Models

The existing quantitative analysis methods cannot predict the fundraising ability, and due to the limitation of the expression ability of traditional analysis methods, there may be problems such as insufficient analysis. Therefore, we put our effort into finding a more effective way to help us analyze the social phenomenon, medical crowdfunding. Recently, machine-learning algorithms have been widely used in various fields such as biology, medical treatment, chemical engineering, and finance. However, there are rare articles about applying machine-learning algorithms to medical crowdfunding for fundraising ability prediction. Although there have been related studies in other crowdfunding fields, such as equity crowdfunding, the data distribution characteristics of medical crowdfunding are obviously different from those in other crowdfunding fields due to the different forms, objects, and goals of crowdfunding. Moreover, there are many kinds of algorithms in machine learning that can find patterns in the training data and map the factors to the target. However, the performance of such algorithms depends heavily on the representation of the collected data and logical inference method. As mentioned in [62], numerous machine-learning algorithms are applicable to a variety of contexts, and the performance of the
algorithms is at varying levels. Thus, it is meaningful to use machine-learning algorithms to analyze data in the field of medical crowdfunding and get a conclusion about the suitable algorithm for specific medical crowdfunding tasks.

In this section, the predictive ability of five representative machine-learning models were compared, i.e., nearest neighbor, linear regression, neural network, decision tree, and an integrated method, to analyze the most suitable machine-learning models for different learning tasks in medical crowdfunding. A brief introduction to each algorithm and the application in medical crowdfunding campaigns are stats below.

K-Nearest Neighbors (KNN) proposed by Cover and Hart [63] in 1967 is a simple widely accepted model that can be used to solve both classification and regression problems for crowdfunding campaigns [54–56]. The main idea of KNN is that a sample should be most similar with \(k\) nearest samples in the data set. The first step is to calculate the Euclidean distance between the test sample and each training sample. Then, we select \(k\) nearest samples after sorting the distance. And the predicting target in this work, fundraising amount, can be calculated by mean value of \(k\) samples. Specifically, the inputs of model KNN are testing sample \(x^*\) and the training data set which contains 11,747 instances with the label (fundraising amount of each instance). Then, the Euclidean distances between \(x^*\) and each training sample are calculated and sorted from small to large. The output of the model is the fundraising amount of test sample \(x^*\), which can be obtained by calculating the mean value of top \(k\) samples.

Linear regression, a traditional method for medical crowdfunding campaigns, is a simple formed and basic regression algorithm, which modeling the linear relationship between our target, fundraising amount \(y\) and the set of 7 factors \(x\). The essential of linear regression is solving the parameters \(\omega\) and \(b\) by minimizing a constructed loss function. Linear regression model employed in this work can be formed as:

\[
\hat{y} = \omega_0 + \omega_1x_1 + \ldots + \omega_mx_m + b = X^TW + b, \tag{1}
\]

where \(\hat{y}\) represents the predicted value of fundraising amount, \(m\) is equal to 7 in this work, because \(\omega\) denotes the weight of the 7 factors, and \(X^T\) is the transpose. The essential of linear regression is solving the parameters \(\omega\) and \(b\) by minimizing a constructed loss function. Also, the mean-squared error is taken as a loss function in this research and defined as:

\[
L(\omega, b) = \frac{1}{n} \sum_{i=1}^{n} (\omega x_i + b - y_i)^2, \tag{2}
\]

where \(n\) is the number of samples which is equal to 11,747 in this work. Thus, the objective function can be written as:

\[
(\omega^*, b^*) = \text{argmin}_{(\omega, b)} \sum_{i=1}^{n} (\omega x_i + b - y_i)^2. \tag{3}
\]

Gradient descent was applied to minimize the distance between the predicted value of fundraising amount and value of training samples. It is considered to be an effective method to approach the minimum value of the objective function, since it can keep updating the parameters (seeking partial derivatives for \(\omega\) and \(b\)) continuously. Specifically, the input of LR in the training stage are samples with description features and the target feature, i.e., fundraising. The weights of each feature would be set randomly at the beginning, the sum of each weighted feature is the output of the model. The weights are updated in each iteration by minimizing the difference between the output and real value of the target feature. In the testing stage, sample \(x^*\) with description features required for fundraising prediction is the input of the well-trained model LR, the fundraising amount for sample \(x^*\) calculated by the sum of weighted description features.

Artificial Neural Networks (ANN), an effective method for medical crowdfunding campaigns [59], abstract the human brain neuron network from the perspective of infor-
formation processing, establishing a simple model and form different networks according to connection methods. Figure 5 illustrates an ANN with four layers, each one containing multiple neurons (nodes). The nodes of adjacent layers are connected with a weighted linear summation and pass values by feedforward. The data set that used in this research contains 7 features and one target, which determines the number of nodes in the input layer is 7. We take ReLU as the nonlinear activation function on each node to transform the result of the linear weighted sum. The activation function is usually smooth and differentiable, such as ReLU, making it possible to train the network based on back-propagation. Specifically, ANN is an extension of LR, which adds the activation function to enhance the expression of the model. In the testing stage, the output and input of ANN are also the predicted fundraising amount and test samples with description features, respectively.

![Artificial Neural Network](image)

**Figure 5.** Artificial Neural Network.

Decision tree is a type of classical machine-learning algorithm able to make predictions for the medical crowdfunding campaigns by generating a tree structure that consists of leaf nodes, internal nodes, and a root node. As one of the representative algorithms of the decision tree, Classification and Regression Trees (CART) [64] proposed by Breiman et al. in 1984 can be used for both classification and regression tasks. In contrast to C4.5 proposed by Quinlan [65] in 1994, CART is essentially a binary division of the feature space (i.e., the decision tree generated by CART is a binary tree). The CART constructed in this work based on 7 factors and a continuous target, fundraising amount. In this regression task we select the best splitting factors for each node by measuring the Least Square Deviation (LSD) of two parts to construct a CART. Specifically, all the samples are stored in the root node at beginning. Samples in the current node split into two parts according to the best splitting factor that minimize the square deviation of both. Finally, the input space will be divided into $M$ regions according to 7 factors. The predicting target, fundraising amount could be obtained by averaging this value of training samples in the same region. The training strategy of CART is to construct a binary tree by selecting the most important description feature before each division of samples. The sample $x^*$ which ask for the target feature, fundraising, is the input in the testing stage, and is would be judged by the description features of each node in the well-trained CART model and assigned to the corresponding child node. The output is the mean value of the fundraising amount that all test samples corresponding to the leaf node where sample $x^*$ located.

Gradient Boosting Decision Tree (GBDT) [66], also known as Multiple Additive Regression Tree (MART), is a representative algorithm based on the Boosting strategy in ensemble learning, which can make predictions for medical crowdfunding campaigns. Chen Tianqi improved GBDT in the Kaggle competition and created the eXtreme Gradient Boosting
(Xgboost) package [67], greatly improving the algorithm’s speed and effect. Here, CART was taken as base learners because it is relatively simple, with low variance and high bias. The model was constructed by adding trees continuously according to the residual between the predicted fundraising amount and the true value of the training sample. A new tree was then included in the model to fit the residual of the previous prediction in each iteration. Predictions are aggregated in an additive manner in which each added model is trained so it will minimize the loss function. The final fundraising amount can be obtained by adding all the scores (each leaf node corresponds to a score) of the well-trained ensemble model with several basic trees. Specifically, Xgboost is an extension of the decision tree, which consists of a set of trees. In the testing stage, the fundraising of a testing sample $x^*$ can be output by summing up the weak results of each tree obtained.

4. Experiments

The experiment consists of two major regression tasks. The first one employs the collected data set to predict the amount of medical crowdfunding fundraising, making a comparison among the results of five different kinds of machine-learning algorithms. Next, samples with medical crowdfunding fundraising rate (the value obtained after dividing the total amount of raised fund by target amount) over 50% are used to predict the fundraising amount.

The experiments were then implemented in Python and ran on a computer with the Windows operating system, an i7-8850H CPU, and 16 GB of RAM (Intel, Santa Clara, CA, USA). One of the steps was to randomly take 80% of the data set as training samples, and the remaining 20% as the testing samples. 10-fold cross-validation was applied in training stage to enhance the generalization ability of the models. In total, five machine-learning algorithms are compared in this work, including CART, KNN, MLP, and Xgboost and the performance of these methods were evaluated based on three metrics: mean absolute error (MAE), mean-squared error (MSE), and R-squared.

4.1. Parameters

The parameters optimized by grid search for each algorithm are mentioned in detail below. In the first task, the Manhattan distance between the target and the 15 nearest points was considered to calculate the regression value in KNN. The maximum depth of the CART is limited to 5 and there are 2 hidden layers in the constructed artificial neural network with 50 and 20 neurons in each one. Here, ReLU is taken as the activation function and the parameter of L2 regularization is 0.001, together with the learning rate that is set as 0.01. The model is trained based on Mean-Squared Loss and optimized with Adam [68], an extension of Stochastic Gradient (SGD). The maximum depth of xgboost is 4, the number of basic regression trees is 100, and the learning rate is 0.35. In task 2, the optimal parameters of some models differ from task 1. For example, the best maximum depth is 3 and 5 for CART and Xgboost. There are 3 hidden layers in ANN with 200, 150, 200 hidden nodes, respectively. The most suitable activate function for task 2 is tanh, a hyperbolic functions.

4.2. Experimental Analysis

The Pearson correlation coefficient is a statistic method that measure linear correlation coefficient and returns a value of between $-1$ and $+1$. The correlations for the factors of medical crowdfunding campaigns, including target, type of raisers, number of retweets, duration, number of donators, number of updates, words of description, and fundraising amount are shown in Table 4. According to the standard proposed by the Political Science Department at Quinnipiac University [69], the correlation between the number of retweets and fundraising, the amount is moderate, and the number of donators and fundraising amounts is strong, indicating that the number of retweets and number of donators are two significant factors influencing the final amount of medical crowdfunding fundraising.
Table 5 reports the results of five algorithms evaluated based on 3 metrics obtained in two different tasks. We visualize the results of two tasks separately into Figures 6 and 7, in which different color represents different machine-learning algorithms, and the length of each bar indicates the Mean absolute error (MAE), Mean-squared error (MSE), and R-squared value of each model. Among all the five algorithms, Xgboost produced the best result across the three performance metrics. In task 1, the performance of all kinds of models, overall, was not satisfactory because of the uneven data distribution and the limited number of attributes that are strongly related to the predicted target. Under such conditions, the performance of Linear regression, CART, ANN, and an ensemble model, Xgboost, are significantly improved compared to the simpler models KNN and LR, with not much difference among these three models. ANN performs best on MAE, while Xgboost performs best on MSE and R-squared. In task 2, we consider samples that fundraising rate is over 50%. Figure 7 illustrates the performance of algorithms improved when we focus on samples that fundraising rate is over 50%. Specifically, MSE decreased and R-squared increased significantly, even though MAE reduced slightly. Xgboost performs best in MSE and MAE while CART performs best in MAE. Hence, the long-tail distribution of data limits the performance of algorithms. We can employ data more effectively when a reasonable data segmentation point is found. The algorithm based on ensemble performs best overall for predicting fundraising performance in medical crowdfunding campaigns.

Table 4. Pearson correlations.

<table>
<thead>
<tr>
<th></th>
<th>Target</th>
<th>Raisers</th>
<th>NumRetweet</th>
<th>Duration</th>
<th>NumDonator</th>
<th>NumUpdate</th>
<th>NumWord</th>
<th>Fundraising</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>1</td>
<td>0.073</td>
<td>−0.017</td>
<td>0.177</td>
<td>0.026</td>
<td>0.108</td>
<td>0.105</td>
<td>0.048</td>
</tr>
<tr>
<td>Raisers</td>
<td>0.073</td>
<td>1</td>
<td>−0.186</td>
<td>0.038</td>
<td>−0.046</td>
<td>0.107</td>
<td>0.047</td>
<td>−0.154</td>
</tr>
<tr>
<td>NumRetweet</td>
<td>−0.017</td>
<td>−0.186</td>
<td>1</td>
<td>−0.103</td>
<td>0.200 *</td>
<td>0.044</td>
<td>−0.132</td>
<td>0.365 **</td>
</tr>
<tr>
<td>Duration</td>
<td>0.177</td>
<td>0.038</td>
<td>−0.103</td>
<td>1</td>
<td>−0.025</td>
<td>0.016</td>
<td>0.107</td>
<td>−0.109</td>
</tr>
<tr>
<td>NumDonator</td>
<td>0.026</td>
<td>−0.046</td>
<td>0.200 *</td>
<td>0.025</td>
<td>1</td>
<td>0.053</td>
<td>0.037</td>
<td>0.455 ***</td>
</tr>
<tr>
<td>NumUpdate</td>
<td>0.108</td>
<td>0.107</td>
<td>0.044</td>
<td>0.016</td>
<td>0.053</td>
<td>1</td>
<td>0.019</td>
<td>0.109</td>
</tr>
<tr>
<td>NumWord</td>
<td>0.105</td>
<td>0.047</td>
<td>−0.132</td>
<td>0.107</td>
<td>0.037</td>
<td>0.019</td>
<td>1</td>
<td>−0.024</td>
</tr>
<tr>
<td>Fundraising</td>
<td>0.048</td>
<td>−0.154</td>
<td>0.365 **</td>
<td>−0.109</td>
<td>0.455 ***</td>
<td>0.109</td>
<td>−0.024</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: *, **, and *** represent correlation is weak, moderate, and strong, respectively.

Table 5. Performance of models in Task 1 and Task 2.

<table>
<thead>
<tr>
<th></th>
<th>Task 1</th>
<th>Task 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>KNN</td>
<td>0.06044</td>
<td>0.01352</td>
</tr>
<tr>
<td>LR</td>
<td>0.07270</td>
<td>0.01566</td>
</tr>
<tr>
<td>ANN</td>
<td>0.04976</td>
<td>0.01057</td>
</tr>
<tr>
<td>CART</td>
<td>0.05147</td>
<td>0.01057</td>
</tr>
<tr>
<td>Xgboost</td>
<td>0.05140</td>
<td>0.01035</td>
</tr>
</tbody>
</table>

Note: The bold highlights the best performed algorithm based on each metric.
5. Conclusions and Future Work

5.1. Discussion Findings

This study first performed preprocessing and exploratory analysis based on Weibo Philanthropy samples and then introduced a series of machine-learning algorithms rarely used in medical crowdfunding before. The 10-fold cross-validation is employed in the training stage, and parameters are optimized by grid search for each algorithm. Indicators mean abstract error, mean-squared error and R-squared are applied to evaluate the performance of algorithms. The experimental results show the performance of Classification and Regression tree, Artificial Neural Network, Xgboost are not much different, outperforming other algorithms, such as K-Nearest Neighbors and Linear Regression. Xgboost constructed based on ensemble performs best among all the algorithms with the smallest mean-squared error and the largest R-squared in both tasks of the experiment. The outcomes demonstrate the potential of the ensemble-based machine-learning algorithms to predict medical crowdfunding project fundraising amounts and provide inspiration for guiding follow-up research and management practices.

5.2. Limitations and Further Research

Considering this is an exploratory study, even though several algorithms were compared and taken into consideration, there are still several limitations that must be acknowledged. First, in order to obtain new data that has not been studied in the past, the data selected in this study was extracted from a single medical crowdfunding platform affiliated to Sina.com which is one of the leading Internet media and services companies for communities worldwide [70,71], however, resulting in limited data. In future research, in order to prove that the conclusion of this study is applicable to more public welfare
crowdfunding platforms and users, we should obtain more data from multiple platforms, such as GoFundme and Tencent Donation.

Also, the selection of attributes needs to be optimized. Since this paper intends to provide new research ideas and the perspective of innovative methods, thus the attribute selection is more based on the actual data and the conclusions of previous empirical studies. However, due to the protection of website information, limited technology and other reasons, the scope of attribute selection still needs to be expanded in future research. It can be a good idea, for example, consider and analyze the influencing factors: the influence of text/image content, the number of potential followers/donors, and the social influence of followers/re-tweeters [6,15].

To improve the performance of the model, data prepossessing and algorithm parameter adjustment are adopted. Rather than stop at the classical model comparison and application, future researchers should consider extending the content of optimization to the construction of the model. For instance, particle swarm optimization-based extreme gradient boosting for predicting the success of medical crowdfunding campaigns might be a good strategy to obtain better performance in medical crowdfunding fundraising predictions.

Author Contributions: Conceptualization: N.P.; methodology, software: X.Z.; validation: N.P. and X.Z.; formal analysis: N.P.; resources: N.P.; writing—original draft preparation: N.P., X.Z.; writing—review and editing: B.N., Y.F.; supervision: Y.F., B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China [Grant No.71971143, 71702111, 71571120]; The Natural Science Foundation of Guangdong Province [Grant No.2020A1515010749]; Major Projects of Universities of Guangdong Provincial Department of Education [Grant No.2019KZDXM030]; Guangdong Province Postgraduate Education Innovation Research Project [Grant No.2019SFKC46]; Postgraduate Innovation Development Fund Project of Shenzhen University [Grant No.315-0000470417].

Acknowledgments: The second author would like to thank her supervisor, Xizhao Wang, for his helpful guidance.

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

References


